EPIGENETIC MECHANISMS IN LUNG CANCER

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Christopher Lee Seiler

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISER: Dr. Natalia Y. Tretyakova

September 2018

Acknowledgments

I would like to offer my gratitude to all who have contributed both directly and indirectly to this dissertation and my education at the University of Minnesota. First and foremost, I would like to thank my adviser, Dr. Natalia Y. Tretyakova, for her continuous guidance and support during my graduate career.

I would also like to thank the many other mentors who have helped shape my scientific career: Peter Villalta and Brock Matter for advancing my skills and knowledge of mass spectrometry techniques employed in this work, Linda von Weymarn and Steve Carmella for instilling a deep appreciation for chromatography, Pramod Upadhyaya and Bin Ma for exemplifying the benefits of thorough discussion, and Delshanee Kotandeniya for inspiring my research into the field of epigenetics. I would like to thank the members of my committee, Courtney Aldrich, Colin Campbell, and Robert Turesky for their guidance and suggestions during my graduate career. I am appreciative of all of the faculty and staff of the Department of Medicinal Chemistry for their influence. I am indebted to all the current and former members of the Tretyakova laboratory and the Masonic Cancer Center, as well as all of the collaborators and contributors to the work presented in this dissertation. I would like to thank Robert Carlson for his constant editorial and technical support.

I would like to thank Alex Baierl and Redmond Fraser for their friendship and support through graduate school. This dissertation would not have been possible with the enthusiasm and support from my parents, family, friends, and fellow graduate students.

Most notably, I would like to thank my wife Sheila Mae Seiler for her support and motivation.

Dedication

This dissertation is dedicated to my wife Sheila Mae Seiler

for her endless love and support.

ABSTRACT

Epigenetic control of gene expression involves covalent reversible modifications of DNA, RNA, and histones which lead to changes in chromatin structure and accessibility. The ability to maintain precise control over gene expression in cells and tissues is critical for ensuring normal cellular development and homeostasis. The most important epigenetic mark of DNA is methylation of cytosine at the C5 position (MeC). This stable epigenetic mark is introduced by *de novo* methyltransferases DNMT3a/b and maintained through cell division by maintenance methyltransferase DNMT1. Ten Eleven Translocation (TET) dioxygenases oxidize 5-methylcytosine (MeC) to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC), a process known to induce DNA demethylation and gene reactivation. A precise balance of DNA methylation and demethylation is important for establishing tissue specific gene expression patterns, maintaining cell identity, and guiding development. However, inflammation and exposure to exogenous agents can lead to changes in DNA methylation patterns or "epimutations" which together with genetic mutations can lead to the development of cancer.

Chapter I of this Thesis provides an overview of the major mechanisms of epigenetic regulation including epigenetic marks of DNA, non-coding RNAs, and histone modifications. Chapter I then describes epigenetic dysregulation in cancer and other human diseases. We then go on to describe how epigenetic changes in DNA can be detected and quantified using antibodies and mass spectrometry-based approaches. After considering global quantitation of epigenetic DNA modifications, we discuss the methods available for mapping epigenetic modifications along the genome.

In Chapter II of this thesis, the effects of C5-cytosine substituents with increased steric bulk on catalytic activity of maintenance DNA methyltransferase (DNMT1) were examined. This protein specifically recognizes 5-methylcytosine (MeC) bases at hemimethylated CG sites in DNA and conducts maintenance methylation. Maintenance methyltransferase activity was the highest towards DNA containing the natural DNMMT1 substrate, MeC. The enzyme was capable of performing maintenance methylation when 5-ethyl-dC was the substrate, while the more rigid and bulky C5-alkyl substituents such as 5-vinyl- dC, and 5-propyl-dC could not direct maintenance methylation.

In Chapter III, we investigated the kinetics of maintenance DNA methylation towards DNA duplexes containing oxidized forms of MeC (hmC, fC, and caC). We also employed a molecular dynamics simulation of the enzyme with the DNA to understand the interactions of oxidized forms of MeC with the DNMT1 enzyme. We found that methyl transfer rates were reduced when MeC was oxidized to hmC, fC, and caC, consistent with the model that Tet mediated oxidation contributes to passive DNA demethylation.

In Chapters IV and V, we investigated inflammation-mediated epigenetic changes in the lung using A/J mouse model of smoking induced lung cancer. In collaboration with NuGEN (Santa Carlos, CA), we developed a novel reduced representation bisulfite sequencing (RRBS) methodology to map both MeC and hmC genome-wide. Our studies provide evidence that inflammation of the lung induces both global and site-specific epigenetic changes in DNA methylation and hydroxymethylation, alters global histone acetylation, and deregulates gene expression. These studies also provide evidence that

exposure to cigarette smoke can alter site-specific DNA methylation and hydroxymethylation of genes that are associated with the cancer phenotype.

The final chapter of this dissertation (Chapter VI) employs affinity proteomics to identify protein readers of epigenetic marks of DNA in the lung. DNA duplexes functionalized with C, MeC, hmC, fC, and caC were attached to solid support and incubated with nuclear protein extracts from human bronchial epithelial cells (HBEC). Proteins specifically recognizing DNA epigenetic marks were identified using Orbitrap Velos mass spectrometer and quantified using 8-plex TMT tags. This chapter details the development of a method for carrying out the affinity proteomics experiments, including solid phase synthesis of DNA targets, peptide tagging, sample clean-up, fractionation, and nanoHPLC-ESI^{+}-MS^{2} based methodology for protein identification and quantification.

Overall, during the course of the studies described in this thesis, we have investigated the specificity and kinetics of human maintenance DNA methyltransferase (DNMT1), employed animal models to characterize epigenetic changes in the lung caused by inflammation and exposure to cigarette smoke, and examined novel mechanisms of epigenetic regulation at oxidized forms of MeC. Overall, this work contributes to current understanding of epigenetic regulation in normal cells and epigenetic deregulation in cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SCHEMES

LIST OF ABBREVIATIONS

2D-LC	two dimensional HPLC
4mC	4-methyl-2'-deoxycytidine
5-ethyl-dC	5-ethyl-2'-deoxycytidine
5-I-dC	5-iodo-2'-deoxycytidine
5-propyl-dC	5-propyl-2'-deoxycytidine
5-vinyl-dC	5-vinyl-2'-deoxycytidine
6mA-RE-Seq	N6MedA restriction enzyme sequencing
ACN	acetonitrile
AGC	automatic gain control
ARP	aldehyde reactive probe
BER	base excision repair
BRCA1	breast cancer type 1 susceptibility protein
BSA	bovine serum albumin
caC	5-carboxyl-2'-deoxycytidine
caMAB-Seq	carboxylcytosine modification assisted bisulfite sequencing
CD	circular dichroism
CDH1	cadherin 1
CDH13	cadherin 13
CDX2	homeobox protein CDX-2
CGIs	CpG islands
ChIP-Seq	chromatin immunoprecipitation sequencing

CMS	cytosine-5-methylsufonate
COPD chronic	obstructive pulmonary disease
Co-REST	corepressor of REST
CpG	cytosine phosphate guanosine
DAPK1	death associated protein kinase 1
DAVID	database for annotation, visualization and integrated discovery
DCM	dichloromethane
DIPEA	diisopropyl ethyl amine
DMF	dimethylformamide
DMT	dimethoxytrityl
DMTrCl	dimethoxytrityl chloride
DNase I	deoxyribonuclease I
DNMT	DNA methyltransferase
DTT	dithiothreitol
EGFR	epidermal growth factor receptor
ELISA	enzyme-linked immunosorbent assay
EMSA	electrophoretic mobility shift assay
Ep-CAM	epithelial cell adhesion molecule precursor
EZH2	histone lysine N-methyltransferase
FASP	filter aided sample preparation
fC	5-formyl-2'-deoxycytidine
fCAB-Seq	formylcytosine chemical assisted bisulfite sequencing

GLIB-seq	glucosylation, periodate oxidation, biotinylation sequencing
HAT	histone acetyl transferases
HBEC	human bronchial epithelial cells
HDAC	histone deacetylases
HKMT	histone lysine methyltransferases
hmC	5-hydroxymethyl-2'-deoxycytidine
hMLH1	DNA mismatch repair protein Mlh1
HMT	histone methyl transferases
HOXA1	homeobox protein Hox-A1
HPLC	high performance liquid chromatography
HPLC-ESI-MS/MS	high performance liquid chromatography electrospray ionization tandem mass spectrometry
HPLC-MS/MS	high performance liquid chromatography tandem mass spectrometry
IgG	immunoglobulin g
IP	immunoprecipitation
IPA	ingenuity pathway analysis
K_m	Michaelis constant
LC-MS/MS	liquid chromatography tandem mass spectrometry
LPS	lipopolysaccharide
LSD	histone lysine demethylase
MAB-Seq	modification assisted bisulfite sequencing

Mbd2	methyl binding domain 2
MD	molecular dynamics
MeC	5-methyl-2'-deoxycytidine
MeDIP	methylated DNA immunoprecipitation
MeOH	methanol
MGMT	methylguanine methyltransferase
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MSO	methylation-specific oligonucleotide
N6MedA	N^6-methyl-2'-deoxyadenosine
NGS	next generation sequencing
NNK	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
NSCLC	non-small cell lung cancer
OPCML	opioid-binding protein/cell adhesion molecule
oxBS-Seq	oxidative bisulfite sequencing
oxiMeC	oxidized forms of MeC
p16	Cyclin-dependent kinase inhibitor 2A
p27	Cyclin-dependent kinase inhibitor 1B
PD-1	programmed cell death protein 1
PDE I	phosphodiesterase I
PDE II	phosphodiesterase II
RAR-B	retinoic acid receptor beta

Rassf1a	ras associated domain-containing protein 1
redBS-Seq	reduced bisulfite sequencing
REST	RE1 silencing transcription factor
RRBS	reduced representation bisulfite sequencing
SAM	S-adenosylmethionine
SILAC	stable isotope labeling of amino acids in cell culture
SMRT	single molecule real-time
SNPs	single nucleotide polymorphisms
SPE	solid phase extraction
Srap1	SOS associated peptidase 1
SRM	selected reaction monitoring
STAGE	stop and go extraction
TAB	TET-assisted bisulfite
TAB-Seq	TET-assisted bisulfite sequencing
TBE	tris borate EDTA buffer
TBS	*t*-butyldimethylsilyl
TBSCl	*t*-butyldimethylsilyl chloride
TDG	thymine DNA glycosylase
TEA	triethylamine
TET	ten eleven translocation protein
Thy28	thymocyte nuclear protein 1
TLC	thin layer chromatography

TMS	trimethylsilyl
TMT	tandem mass tags
TRD	target recognition domain
TSN	tobacco specific nitrosamine
TSQ	triple stage quadrupole
Uhrf1	E3 ubiquitin-protein ligase Uhrf1
UV	ultraviolet
UV-VWD	ultraviolet variable wavelength detector
V_{max}	maximum velocity
WGBS	whole genome bisulfite sequencing

I. LITERATURE REVIEW

Adapted in part from:

Seiler, C.L.; Fernandez, J.; Han, Q.; and Tretyakova, N.Y.; Experimental methodologies for detection and mapping of epigenetic DNA marks, *Springer Chapter*, submitted March 19, 2018.

This thesis chapter is an expansion of a book chapter submitted to the Springer publishing group, which was a collaborative effort between Christopher Seiler, Jenna Fernandez, and Qiyuan Han, under the direction of Prof. Natalia Tretyakova. Christopher Seiler wrote sections 1.1, 1.2, and 1.7. Christopher Seiler, Jenna Fernandez, and Qiyuan Han wrote sections 1.3, 1.4, and 1.6. Qiyuan Han wrote section 1.5. Christopher Seiler and Jenna Fernandez edited the text and prepared the original figures.

1.1 Epigenetic regulation of gene expression

1.1.1 What is epigenetics?

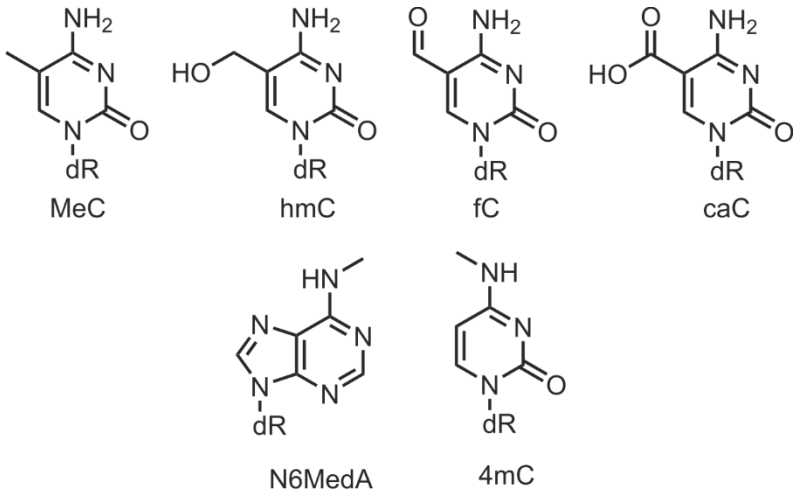
Generally, epigenetics is defined as changes in gene expression that do not result from changes in the gene sequence.[1] Epigenetic regulation can be achieved through reversible covalent modifications of DNA, RNA, and histones.[1] DNA epigenetic modifications are primarily found as cytosine methylation, its oxidized forms, and may also employ methylation of adenine (see section 1.1.2 for more information).[1] Epigenetic modifications of histones include lysine acetylation, which can reduce the charge on histones, and methylation, which can occur as mono-, di-, or tri-methylated forms (see section 1.1.3 for more information).[1] Epigenetic regulation of RNA occurs through individual RNA species such as miRNA as well as covalent modification of mRNA transcripts (see section 1.1.4 for more information).[1] Epigenetic changes are critical in cellular development and differentiation, as well as for normal cellular processes.[1] Epigenetic deregulation is increasingly recognized as a critical event in the development of cancer and other diseases including asthma,[2] autism, and diabetic cardiomyopathy.[3-4]

1.1.2 DNA epigenetic marks

Epigenetic control on DNA primarily presents as methylation of C5-position of cytosine.[5] The human genome contains 3 billion base pairs.[6] Of these, nearly 27 million cytosines exist in the CpG sequence context, making them potential targets for DNA methyltransferase (DNMT) activity.[7] Cytosine methylation can also be found in CHH and

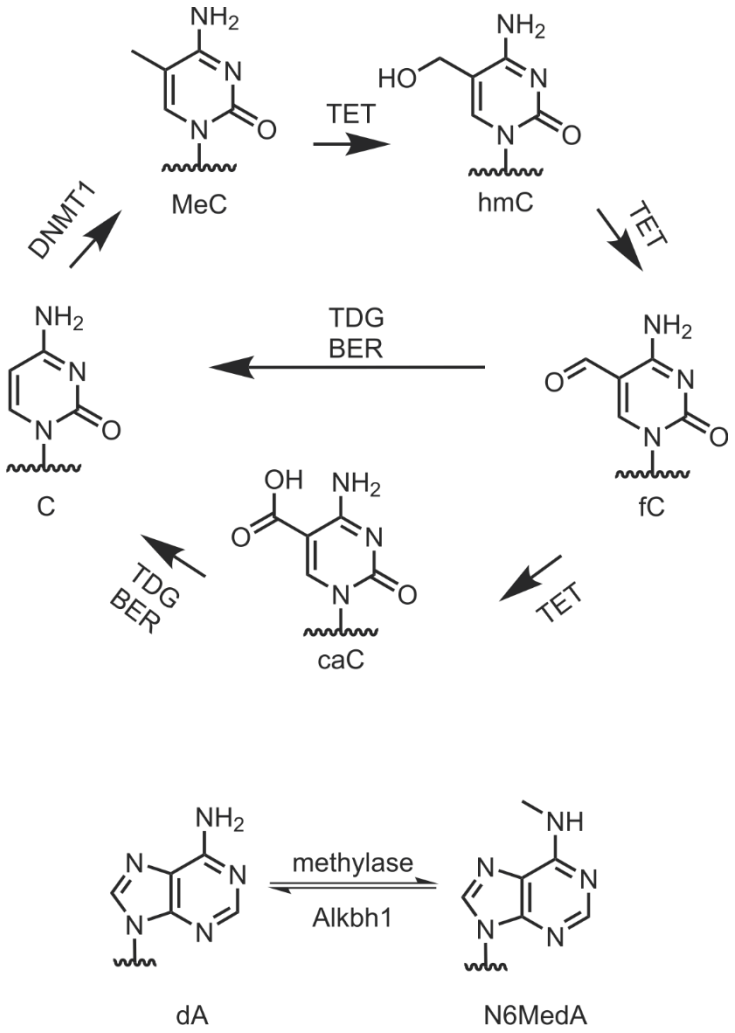
CXG contexts, where H is A or T. It is currently unknown what enzymes in mammals may be responsible for methylating cytosine in these contexts. 5-Methylcytosine (MeC in **Figure 1.1**) is a key epigenetic modification of DNA with important roles in gene regulation, genetic imprinting, X-chromosome inactivation, and genome stabilization.[8] Methylated CpG sites are recognized by methyl-CpG binding proteins, leading to the recruitment of histone deacetylases, chromatin remodeling, and in most cases, reduced levels of gene expression.[9-10] DNMTs introduce MeC into the genome by catalyzing the transfer of a methyl group from S-adenosylmethionine (SAM) to the C-5 position of cytosine in DNA.[10-11] DNMT3 A/B are responsible for *de novo* DNA methylation, a process mediated by the DNMT3L scaffolding protein.[12] Once established, DNA methylation patterns are faithfully maintained by specialized DNA methyltransferase, DNMT1, which recognizes hemi-methylated CpG sites and methylates the newly synthesized strand.[10] This preserves cell identity and maintains tissue specific gene expression patterns following semi-conservative DNA replication.

Figure 1.1. Structures of epigenetic modifications of DNA including 5-methyl-2′-deoxycytidine (MeC), 5-hydroxymethyl-2′-deoxycytidine (hmC), 5-formyl-2′-deoxycytidine (fC), 5-carboxyl-2′-deoxycytidine (caC), N^6-methyl-2′-deoxyadenosine (N6MedA), and 4-methyl-2′-deoxycytidine (4mC).



Recent studies have shown that cytosine methylation marks are reversible. Ten eleven translocation family of proteins (TET 1-3) iteratively oxidize MeC to generate 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) (**Figure 1.1 and Figure 1.2**).[13-16] TET enzymes utilize active site non-heme iron and α-ketoglutarate to split molecular oxygen and sequentially oxidize MeC.[13, 17-18] During the oxidation of MeC to hmC, fC, and caC, hydrogen atom abstraction from the C5 methyl group is the rate limiting factor.[17] In the reaction mechanism, a hydrogen atom needs to be directed toward the active site iron for the oxidation reaction to proceed.[17-18] The C-5 methyl group of MeC is able to rotate and position its hydrogen atoms efficiently in the active site of the protein. Once MeC is oxidized to hmC and fC, the abstraction of hydrogen atoms from C5 becomes progressively more difficult, slowing the enzymatic activity of Tet.[17] When fC is present in the active site, the C5-carbonyl oxygen and N^4-exocyclic amine form an intramolecular hydrogen bond, which keeps the remaining C5-hydrogens 4.98 Å away from the active site iron.[17] The hydroxymethyl group of hmC can also form an intramolecular hydrogen bond with the exocyclic amine of the pyrimidine. However, this nucleoside can exist in two conformations in one of which the C5-hydrogen atom is oriented toward the catalytic iron, allowing for oxidation to occur, at a slower rate.[17]

Figure 1.2. Enzymatic pathways that generate MeC, hmC, fC, caC, and N6MedA in genomic DNA. Cytosine is methylated by DNA methyltransferase in the presence of S-adenosylmethionine. MeC can then be iteratively oxidized by the TET family of enzymes to produce hmC, fC, and caC. N6MedA is introduced by a yet undiscovered methylase and can be demethylated by Alkbh1.



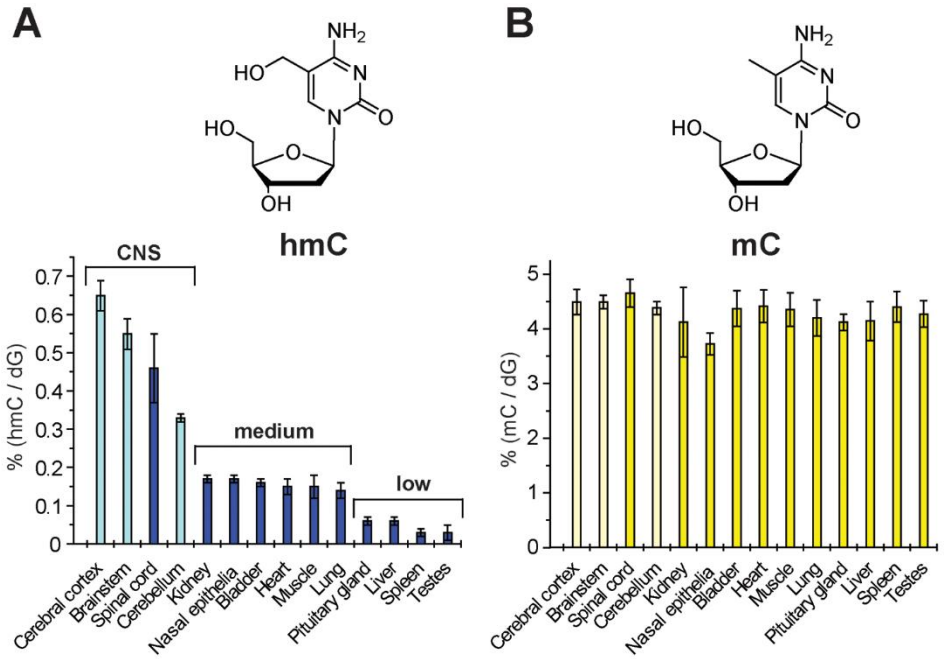
TET-mediated oxidation of MeC to hmC, fC, and caC leads to passive DNA demethylation since DNMT1 activity is reduced in the presence of oxidized forms of MeC .[19-20] Both fC and caC can also be removed via thymine DNA glycosylase (TDG) and replaced with C via the base excision repair pathway, leading to active DNA demethylation.[16] Oxidized forms of MeC can also be enzymatically removed by Srap1endonuclease (SOS associated peptidase 1).[21]

In addition to their roles in DNA demethylation, oxidized forms of MeC have been proposed to possess their own epigenetic functions by directly participating in epigenetic signaling.[22-25] Mass spectrometry based proteomics studies have identified a number of proteins that specifically interact with MeC, hmC, fC, and caC, suggesting that each of these epigenetic marks is recognized by a unique set of "readers".[26-27] For example, Mbd2 specifically binds MeC, Thy28 specifically binds hmC, while Uhrf1 preferentially binds MeC and hmC, and Srap1 binds hmC, fC, and caC.[21, 26] Genome-wide distributions of MeC, hmC, fC, and caC bases further suggest that they play different roles in gene regulation.[28] For instance, hmC is preferentially enriched at distal regulatory elements such as enhancers in mouse and human embryonic stem cells, whereas fC or caC tend to accumulate at poised enhancers and promoters, respectively.[28-29]

MeC and its oxidized forms are present in all mammalian tissues examined, but in different amounts. MeC makes up 4.30 ± 0.22% of dG.[30] hmC levels vary between 0.03% and 0.69% of dG depending on tissue type (**Figure 1.3**).[30] In most tissues, fC and caC are present in much lower amounts, with fC concentrations approximately three orders of

magnitude lower than the levels of hmC, and the levels of caC approximately 10-fold lower than the amounts of fC.[31]

Figure 1.3. Tissue specific amounts of MeC and hmC determined by mass spectrometry. **A.** hmC abundance is tissue-dependent, with the highest levels observed in tissues of the central nervous system. The lowest levels of hmC are observed in the pituitary gland, liver, spleen, and testes. This roughly correlates with cell proliferation rates, with higher levels of hmC found in tissues with the lowest proliferation rates. **B.** Values of MeC expressed as percentage of dG. Levels of MeC appear to stay relatively consistent across all tissue types. Adapted from Globisch D et al. (2010) Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS One. Dec 23;5(12):e15367. doi: 10.1371/journal.pone.0015367. Used in accordance with CC BY license (http://creativecommons.org/licenses/by/4.0/)." [30]



Recently, several additional modifications of DNA were discovered including N^6-methyl-2′-deoxyadenosine (N6MedA) and N^4-methyl-2′-deoxycytidine (4mC) (**Figure 1.1**), which may play a role in epigenetic regulation. These DNA modifications have long been known to exist in prokaryotes, where N6mdA and 4mC are the most prevalent DNA modifications and are primarily used for distinguishing host DNA from foreign pathogenic DNA.[32-33] First indirect evidence for the presence of N6MedA in mammals dates to 1983.[34] Although subsequent studies by Ratel et al. were unable to confirm the presence of N6MedA in mammalian genomes,[35] more recent reports confirmed the observation of N6MedA in mouse and human genomes.[33, 36] N6MedA is widely distributed across the genome, although depleted in exonic regions.[33] Furthermore, N6MedA is preferentially found at TAGG sites and has been implicated in gene repression.[33, 36] Recently, N6MedA in mouse brain has been shown to be mediated by environmental stress, although the biological significance of this finding remains to be elucidated.[37] Many biological features of N6MedA in mammals require further study, including the nature of N6MedA writers, erasers, and readers.[33]

N^4-methyl-2′-deoxycytidine (4mC) has been observed in DNA of thermophilic bacteria and many bacteria mesophiles.[38] In bacteria, 4mC is generated by N-4 cytosine specific DNA methyltransferases.[39] 4mC is a part of a bacterial restriction modification system used to defend against foreign DNA, with several 4mC methyltransferases and 4mC-sensitive restriction endonucleases identified in various bacterial strains.[40] However,

additional biological functions of 4mC have not been fully elucidated, and its possible function in eukaryotes remains to be established.

DNA epigenetic marks have distinct relationships with the levels of gene expression as indicated by the varied genomic locations determined by DNA-Sequencing (See section 1.4 for more information). MeC is consistently increased in promoter CpG islands of inactive genes and in inactive promoter regions with low CpG density.[41] hmC is enriched in TSS and gene bodies and is associated with active gene transcription, a reversal of MeC repression.[42] fC and caC are associated with active gene transcription.[43-44] However, fC and caC are associated with reduced rates of RNA polymerase II catalyzed transcription and substrate specificity.[45]

1.1.3 Histone epigenetic marks

Mammalian DNA is packaged in the nucleus, where it exists in complex with histone proteins to form chromatin.[1] Approximately 147 base pairs of DNA is wrapped around a histone octamer comprised of two copies of each histone H2A, H2B, H3, and H4 or their variants.[46] The presence of epigenetic marks on histones has been known since 1964.[47] These marks occur on the highly basic lysine rich tails of the histones and alter the chromatin structure and interacting proteins.[46] Histone tail lysines are modified with acetyl and methyl groups with the methylation occurring as a mono-, di-, or tri-methylated.[48] Generally, histone acetylation removes the positive charge on histone tails by converting a primary amine group ($-NH_2$) to an acetamide ($-NH-C(O)-CH_3$). This change reduces

charge-charge interactions of histones with the negatively charged phosphodiester backbone of DNA, allowing for chromatin opening and active transcription.[49] Histone epigenetic marks are introduced by a group of enzymes termed writers (histone acetyl transferases, HAT, histone methyltransferases, HMT), interpreted by readers for acetylation (bromodomains) and methylation (chromodomains), and removed by erasers (histone deacetylase, HDAC, histone lysine demethylase, LSD).[46, 50]

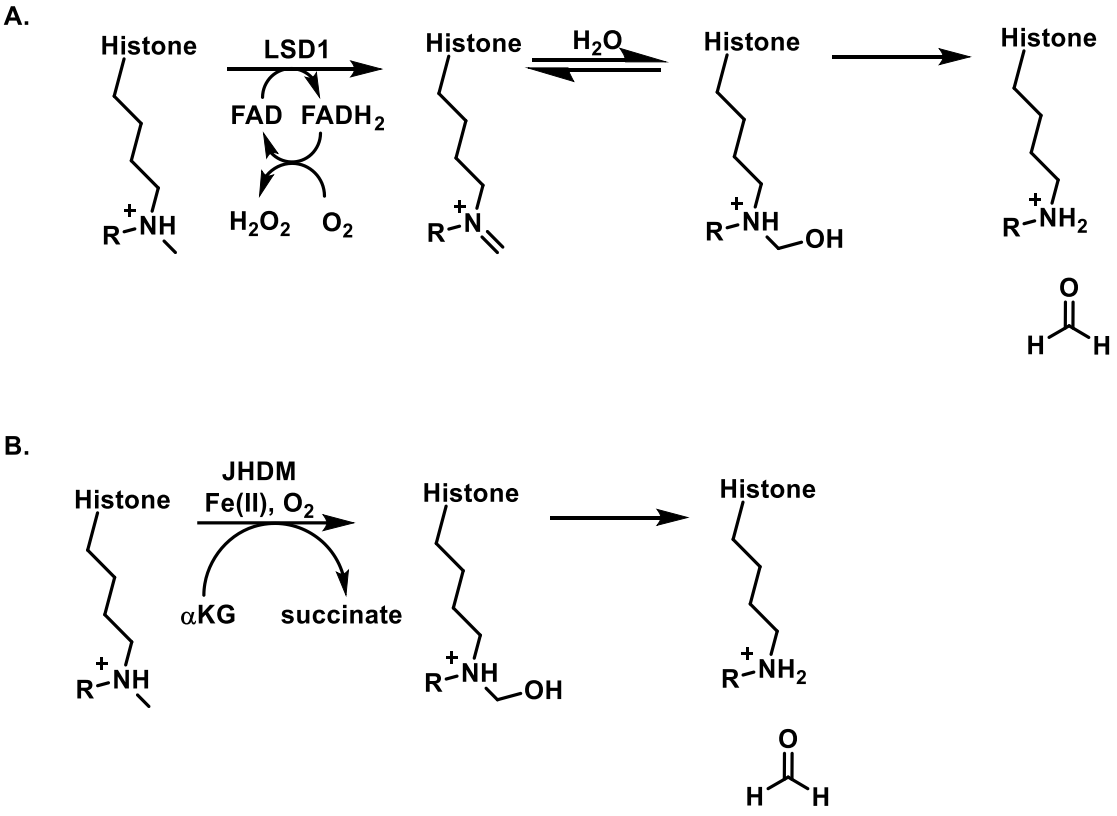
Addition of acetyl groups to histones by HATs utilizes the acetyl group from acetyl-CoA and requires a catalytic glutamate residue in the enzyme active site.[51] The effect of histone acetylation of H3K9, H3K14, and H4K12 on chromatin structure has been investigated using chromatin immunoprecipitation sequencing (ChIP-Seq).[52] Histone acetylation is correlated with actively transcribed genes.[46] In particular, H3K9 acetylation is associated with transcriptional start sites. Similarly, acetylation of H2AK5, H3K14, and H4K12 is common in gene bodies of highly expressed genes.[48] Many HATs are present in multiprotein complexes which influence and direct their activity.[53] HDACs oppose the activity of HATs by removing the acetyl groups from lysine, restoring the positive charge on histones and generally acting as transcriptional repressors.[46, 48] Similarly to HATs, HDACs can also function in protein complexes containing additional HDACs, which modulates the activity and specificity of the HDACs in the complex.[46, 48, 54] Bromodomain family of protein readers is implicated in many regulatory pathways, including cell signaling and chromatin remodeling.[46] Examples of bromodomain proteins include BRDT, BRD2, BRD3, and BRD4.[46]

Histone methylation is catalyzed by histone lysine methyl transferases (HKMT).[48] These proteins function by transferring the methyl group from SAM to an ε-amino group of lysine. HKMT proteins tend to be quite specific for their substrates and in some cases can catalyze the formation of di- and tri-methylated lysine.[48, 55] The ability of HKMT proteins to catalyze the formation of trimethylated lysine is typically controlled by the presence of phenylalanine in their active site, and replacement of the Phe with Tyr prevents formation of trimethylated lysine.[55-56] Histone lysine methylation, e.g. H3K27, H3K9, H3K4, H4K20, H3K79, and H2BK5, affects the levels of gene expression.[57] Interestingly, the exact methylation state of the lysine is very important: H3K27 trimethylation being associated with repression, while H3K27 monomethylation associated with activation.[58-59] Additionally, repression of transcription by trimethylation of H3K27 is associated with an increase in CpG methylation.[60]

Histone demethylation is catalyzed by lysine specific demethylases (LSD).[48, 57] LSD1 recognizes mono- and di-methylated H3K4 in free histones, however, LSD1 can also recognize nucleosomes when it is a part of the Co-REST system.[48, 61] Co-REST is a protein complex that functions as a corepressor of the REST (RE1 silencing transcription factor).[62] LSD1 is part of the KDM1 family of lysine specific demethylases.[63] KDM1 proteins are able to demethylate mono- and di-methylated lysines via a catalytic mechanism that includes the formation of an imine intermediate (**Scheme 1.1A**).[48, 64-65] The imine is subsequently hydrolyzed, giving rise to free Lys and releasing formaldehyde.[64] The need to form an imine intermediate prevents KDM1 family demethylases from

demethylating trimethylated lysines. The second family of LSDs has a mechanism distinct from that of LSD1 and instead contain a Jumonji C domain that performs an oxidative demethylation reaction to remove methyl groups from lysine (**Scheme 1.1B**).[66] The Jumonji C domain uses Fe(II) and α-ketoglutarate to oxidize the methyl group to a hydroxymethyl which spontaneously breaks down to a free amine and liberates formaldehyde.[48, 66]

Scheme 1.1. Histone lysine demethylation by LSD1 (**A**) and Jumonji C protein JHDM (**B**).



1.1.4 RNA epigenetic marks

Cellular RNA exists in multiple forms including mRNA, miRNA, lncRNA, snoRNA, snRNA, rRNA and tRNA.[67] Several RNA species are involved in the epigenetic regulation of gene expression. mRNA is heavily modified post transcription with over 160 known modifications.[68-69] These post-transcriptional modifications can act to alter RNA splicing, mRNA translation rates, and the longevity of mRNA in the cell.[68-69] Further control is imposed by micro RNA species (miRNA). miRNA (an RNA species about 22 nucleotides in length) act by binding to the 3′UTR of target mRNA species to inhibit translation and to increase its degradation.[70-71] Many mRNA contain multiple miRNA binding sites, allowing for a complex regulatory network.[67] miRNA are widespread and conserved regulatory molecules, although some are tissue specific.[72-73] In addition, miRNA have been shown to control a diverse group of processes including pluripotency, epithelial-mesenchymal transition, and metastasis among others.[73-75] Recently, long noncoding RNA (lncRNA, a non-protein-coding RNA transcript greater than 200 nucleotides) emerged as a component of cellular differentiation.[4, 67, 76-77] Epigenetic regulation by lncRNA is achieved through multiple pathways including recruitment of chromatin modifying complexes, targeting DNA for methylation, and enhancing the expression of neighboring cognate genes.[76, 78-79] These effects are mediated through lncRNA in complex with chromatin modifying complexes such as HMTs.[80-81]

1.2 Epigenetic deregulation and human disease

1.2.1 Role of epigenetic deregulation in disease

Epigenetic control of gene expression allows for a regulated expression of proteins and signaling molecules necessary for cells and tissues to perform their unique functions. Genetic mutations found in tumors can alter a protein's sequence that affects function and dysregulated epigenetic signaling can affect the expression and levels of proteins in the cell. Epigenetic changes, or "epimutations", are increasingly recognized as critical events in the development of cancer and other diseases.[3] Specifically in cancer, changes in DNA methylation patterns can lead to silencing of tumor suppressor genes and activation of protooncogenes, which represents an early critical event in tumor development.[3] Aberrant cytosine methylation is implicated in many additional human diseases including asthma, autism, and amyotrophic lateral sclerosis.[2, 82-83] Drugs that inhibit DNA methylation have been successful for treating many types of cancer.[84] DNMT inhibitors such as 5-azacytidine, 5-aza-2′-deoxycytidine (decitabine) and pyrimidin-2-one ribonucleoside (zebularine) are used as single agents and in combination with classical chemotherapeutic agents in treatment of various tumors including myelodysplastic syndromes.[85-89]

DNA hydroxymethylation is also implicated in many diseases. Elevated levels of hmC are observed in patients with systemic lupus erythematosus,[90] autism,[82] and bronchial asthma.[91] In contrast, the levels of hmC are substantially lowered in human cancers, with a 50% to 90% reduction of hmC in breast cancer, liver cancer, lung cancer, and other tumor

types.[92] There are currently no drugs targeting the TET enzymes available on the market and no TET specific inhibitors.

Epigenetic modifications of histones have also been assessed for their role in human disease. HDAC inhibitors have been used as anticancer agents in combination with other therapeutics, this includes FDA approved drugs Vorinostat and Panobinostat.[93-94] Inhibitors of the histone acetylation readers, bromodomain proteins, have been explored as possible therapeutics for a variety of conditions including cancer, inflammatory and infectious diseases.[48] The most well-known bromodomain inhibitor, JQ1, has been shown to have promising anticancer effects.[95]

Noncoding RNA, especially miRNA, has been associated with a variety of diseases including cancer, myocardial infarction, and autoimmune disease.[96-97] RNA regulation can be achieved by transcribing pseudogenes, which are chromosomal regions that are an imperfect copy of a functional gene but will often carry the same miRNA binding sites.[98] The transcription of a pseudogene's mRNA in the cell function as a miRNA decoy for the cognate gene and can reverse the repression of those cognate genes by intercepting the repressive miRNAs.[98] miRNA targets have been suggested as potential therapeutic targets using strategies such as introduction of mature miRNA duplexes to increase levels.[97] These approaches have similar difficulties to gene therapy approaches such as selective targeting and transient expression, although some easily accessible sites like the lungs and the eye have been successful.[99-100]

1.2.2 Epigenetic changes in lung cancer

Lung cancer is anticipated to kill over 150,000 Americans this year and is responsible for 30% of cancer deaths worldwide.[101] Cigarette smoke contains over 60 known carcinogens including tobacco specific nitrosamines and butadiene as well as co-carcinogens such as the bacterial endotoxin lipopolysaccharide (LPS).[102-105] These carcinogens and co-carcinogens can induce genetic and epigenetic alterations.[104, 106] Among epigenetic changes detected in smoking-induced lung tumors, DNA methylation, histone acetylation, and micro RNA expression are often dysregulated, leading to silencing of tumor suppressor genes and activation of oncogenes such as *Kras*.[104] In particular, methylation of key genes such as MGMT, Rassf1a, DAPK1, p16, hMLH1 correlates with malignant phenotype.[107-108] Rassf1 promoter hypermethylation has been shown to be present in invasive cancers.[109] For lung adenocarcinoma, Tsou et al. identified a panel of 5 genes (p16, EX2, CDX2, HOXA1, and OPCML) whose methylation state could selectively distinguish cancerous and non-cancerous tissue.[110] p16, EX2, CDX2, HOXA1, and OPCML are identified as having a high frequency of DNA methylation and a significant increase in methylation in tumors compared to adjacent normal tissue.[110] A total of 28 loci were tested in matched tumor and adjacent normal tissue to find the most predictive biomarkers.[110]

In addition to changes in DNA methylation, the levels of histone modifying enzymes are altered in lung cancer.[111] This includes the methyltransferase EZH2, which methylates H3K27 and is associated with silencing CDH1, p27, and BRCA1.[111-113] HMT

G9a is also capable of methylating H3K37 and promotes lung cancer invasion and metastasis by silencing Ep-CAM.[114-115] Histone lysine acetylation marks are also deregulated in lung cancer, especially hypoacetylation of H3K18.[116-117]

Some miRNAs can act as oncosuppressors, while others, such as miR-21 or miR-663, act as oncogenes and are overexpressed in cancer.[118-119] miR-21 in particular has been is frequently amplified in non-small cell lung cancer (NSCLC).[120] miRNA-663 reduces apoptosis by limiting the permeability of the mitochondria outer membrane by inducing p53 translocation to the mitochondria.[119] miR-663 depletion was enough to cause cell death in NSCLC *in vivo*.[119]

Advances in the treatment of lung cancer over the past 25 years have increased the 5-year survival rates from 12.3% (1975-77) to 16.3% (2001-7).[121] Many new treatments have been approved by the FDA, including angiogenesis inhibitors, antimetabolites, antimicrotubular, EGFR inhibitors, kinase inhibitors, PD-1 blockers, and a photosensitizing agent. Unfortunately, initial clinical trials have failed to achieve a significant response in lung cancer patients with epigenetic drugs.[122-126] New agents targeting additional epigenetic pathways are needed to help improve lung cancer treatment.

1.2.3 Inflammation and increased cancer risk

Tissue inflammation is a necessary process employed in the innate immune system and during wound repair.[127] However, chronic inflammation is strongly associated with the pathogenesis of cancer.[128-129] Specifically, chronic inflammation in the lung such as in

chronic obstructive pulmonary disease (COPD) increases the risk for lung cancer, while colitis and Crohn's disease are associated with colon cancer.[130] COPD patients are at 5-fold higher risk for lung cancer, while Crohn's disease can increase colon cancer risk between 4 and 20-fold.[131-132] Chronic inflammation caused by *H. pylori* infection in the stomach is associated with gastric cancer, while HPV can cause cervical cancer, and hepatitis viruses - liver cancer.[129]

1.2.4 Animal models of lung cancer initiation

Various animal models have been used for studying lung cancer initiation, with the two most common being murine systems.[133] Rat models of smoking induced lung cancer have the advantage of little to no spontaneous background tumor formation. Female rats exposed to mainstream cigarette smoke had a significant increase in lung tumorigenesis.[134] One of the key animal models of smoking-induced lung cancer is the A/J mouse.[135] A/J mice are susceptible to the formation of spontaneous and induced lung tumors and have been widely used to observe lung tumorigenesis following exposure to environmental tobacco smoke.[133, 135] Multiple studies of A/J mice exposed to mainstream smoke for 5 months with 4 months post-exposure have shown a positive correlation of exposure with lung tumorigenesis.[136-138] In these smoke exposure studies, whole body exposure to mainstream smoke as opposed to nose only exposure produced positive tumorigenesis results.[136] Life-time exposure to cigarette smoke has also been examined in female B6C3F1 mice. Hutt et al. found a significant increase in the lung tumorigenesis after life-time

exposure (over 900 days) to mainstream cigarette smoke.[139] The exposure of A/J mice to cigarette smoke starting 12 hours after birth for 4 months followed by 5 months post exposure resulted in an increase in the number of lung tumors in exposed mice.[140] This tumorigenic effect was reduced in mice which were simultaneously dosed with anti-inflammatory agents such as acetylsalicylic acid, supporting a critical role of inflammation in smoking-induced lung cancer.[140]

Animal models have also been used to explore the carcinogenicity of individual tobacco carcinogens such as tobacco specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and cocarcinogens such as lipopolysaccharide (LPS).[141-143] Using the A/J mouse model, Melkamu et al. observed that the inflammatory agent LPS can increase both the size and multiplicity of NNK-induced tumors, supporting the hypothesis that lung inflammation promotes tumorigenesis.[141] A further study with lung inflammation by LPS in the A/J mouse found that this effect could be attenuated by dietary supplements of indole-3-carbinol and silibinin.[142] Given individually to A/J mice by intragastric gavage, NNK and benzo[a]pyrene were found to induce lung tumor formation following 8 week long exposures, but the tumorigenicity could be reduced by pre-dosing with chemopreventative agents benzylisothiocyanate and phenethylisothiocyanate.[144]

Overall, previous studies using rodent models have established that lung tumors can be induced by exposures to tobacco smoke and to single tobacco carcinogens such as NNK, and that this effect can be enhanced by inflammation. These investigations have

confirmed that lung tumors can be caused by exposure to cigarette smoke or its components, consistent with epidemiological evidence.

1.3 Methods for detection and quantification of epigenetic modifications within the genome

1.3.1 Initial detection of MeC and its oxidized forms via paper chromatography and thin layer chromatography

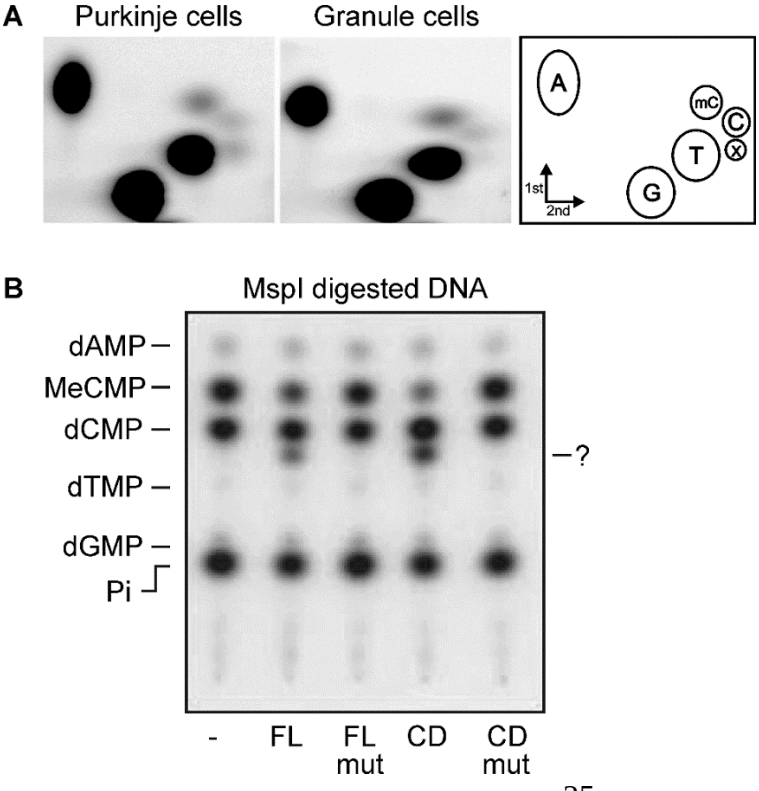
Paper chromatography and thin layer chromatography experiments played a key role in the initial discovery of epigenetically modified DNA bases.[145] In paper chromatography, a compound mixture is spotted on a strip of chromatography paper, which is then hung in a development chamber, and dipped into a solution of volatile alcohols in which to develop the chromatogram. Compounds that are more soluble in the alcohol mobile phase move up the paper strip during development. Paper chromatography was used for the initial detection of MeC in eukaryotic DNA.[145] Calf thymus DNA was hydrolyzed with acid, and the resulting free bases were separated by paper chromatography using butanol. The resulting chromatographic spots were extracted with alcohol and identified by comparing their UV spectra to those of standard nitrogenous bases. Paper chromatography has since been replaced by newer techniques such as thin layer chromatography (TLC).

TLC was instrumental for the initial discovery of hmC in Purkinje cells and cerebellum tissue[146] and the identification of TET family of enzymes responsible for the

oxidation of MeC to hmC.[13] To determine the components of DNA by TLC, the DNA was digested to nucleosides and radiolabeled with ^{32}P, followed by 2D-TLC in which an additional spot was observed on the TLC plate.[146] The new TLC spot, corresponding to hmC, was detected in both cell types, but was twice as abundant in Purkinje cells as compared to granule cells.[146] TLC was also used to identify hmC as the oxidation product of MeC in HEK293 cell DNA.[13] Following transfection of HEK293 cells with a plasmid containing TET1, cellular DNA was isolated and cleaved with MspI. Following radiolabeling with ^{32}P, DNA fragments were enzymatically digested to nucleotides, which were run out on a cellulose TLC plate using a mixture of isobutyric acid, water, and ammonia and imaged using a phosphoimager.[13] The TLC showed a small spot that was only visible when cells had been transfected with TET1 and cleaved with MspI (**Figure 1.4**).[13]

The strength of TLC in studies aiming to identify new epigenetic modifications of DNA lies in its high sensitivity following ^{32}P end labeling of native and modified nucleosides. A key weakness of TLC is the necessity for a secondary confirmation of the identified spot, which is most often done through mass spectrometry. Additionally, TLC is not scalable to a high throughput format and thus is not well suited for large studies.

Figure 1.4. Identification of hmC in brain cells and its formation by TET1 via ^{32}P TLC. **A.** The 2D-TLC of Purkinje cells and Granule cells indicated a new spot marked X, which comigrates with hmC mono phosphate. From Kriaucionis, S., and Heintz, N. (2009) *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science 324, 929-930. Reprinted with permission from AAAS.[146] **B.** DNA was isolated from HEK293 cells after transfection with TET1 plasmid. DNA was digested with MspI, ^{32}P end labeled, and digested to nucleosides. HEK293 DNA revealed an additional nucleotide in the presence of TET1, both full length (FL) and catalytic domain (CD), but not in the mutants of either. From Tahiliani, M. et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930-935.[13] Reprinted with permission from AAAS.



1.3.2 Antibody based detection of epigenetic marks in DNA

Antibodies that specifically recognize epigenetic marks of DNA are commonly used to quantify their global levels. In 1980, Sano et al. developed an antibody against MeC which could recognize methylated sites in purified DNA.[147] In brief, DNA fragments were separated by agarose gel electrophoresis and transferred to nitrocellulose paper. The membrane was incubated in the presence of purified antibody raised in rabbits against MeC, followed by incubation with ^{125}I-labeled goat anti-rabbit IgG. The MeC was then visualized by autoradiography. This methodology was capable of detecting 20 fmol of MeC in immobilized restriction fragments of DNA.[147] A more sensitive MeC antibody, as well as an antibody specific for N6MedA, were developed a few years later and utilized for their detection in human and *Drosophila* DNA.[34] This more sensitive MeC antibody enabled the detection of a minimum of 5 fmol amounts in nanogram quantities of DNA.[34]

Today, antibody-based detection of DNA epigenetic marks is a commonly used technique, and antibodies for many epigenetic DNA modifications are commercially available. These antibodies have frequently been used in an enzyme-linked immunosorbent (ELISA)-type assay (**Figure 1.5**).[148-149] Commercially available ELISA-based kits for global DNA methylation profiling are available from a number of companies including Zymo and EpiGentek. Generally, these kits capture DNA on an ELISA plate. MeC is detected through sequential incubations with a primary antibody against MeC, followed by a secondary antibody linked to luciferase or a similar enzyme, and finally using colorimetric or fluorometric detection reagents.[148] The development of antibodies specific

for hmC and fC allowed for their quantitation using similar commercially available kits (e.g. EpiGentek MethylFlash Global DNA hydroxymethylation (5-hmC) ELISA Easy Kit; Active Motif Global 5-hmC DNA Quantification Kit; EpiGentek MethylFlash 5-Formylcytosine (5-fC) DNA Quantification Kit). These ELISA-based assays are easy to perform, require equipment that is readily available at most research laboratories, and work best for the identification of large changes in global DNA methylation. However, it should be noted that ELISA-based methods are prone to high variability, do not allow for absolute quantitation, and are best suited for detecting large changes in global levels of hmC or fC.

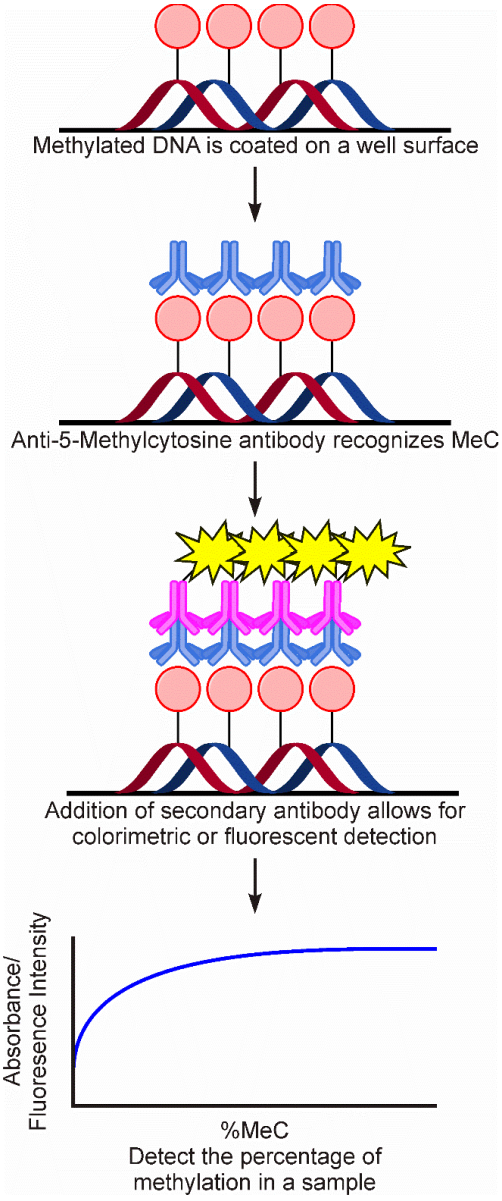
Dot blotting is another simple and rapid antibody-based method to quantify global levels of epigenetic modifications of DNA. In this method, genomic DNA is spotted on a membrane. The membrane is sequentially incubated in blocking buffer, antibody solution, and rinsing buffer. Using an antibody specific for a specific epigenetic modification allows for detection of overall global genomic levels. After chemiluminescence imaging, the intensity of the signal can be used to compare the levels of epigenetic modifications between samples. Generation of antibodies specific for different epigenetic modifications has permitted the quantification of MeC, hmC, fC, caC, and N6MedA in both human and mouse tissues by dot blot.[33, 150-151] As is the case for ELISA-based detection, dot blots do not allow for absolute quantification and suffer from high variability.

Antibodies have also been used to elucidate the biological functions of DNA epigenetic marks by studying their tissue-specific levels and distribution. Global levels of epigenetic DNA marks can be evaluated by staining cells or tissues with a specific

monoclonal antibody.[152] Globisch et al. utilized immunostaining experiments with a commercially available hmC-specific antibody to map the distribution of hmC in cells and tissues. They found that hmC was localized in the cell nuclei and that virtually all cells contained hmC. Furthermore, the use of immunostaining allowed these authors to confirm that the highest levels of hmC are found in the brain, followed by kidney and nasal epithelia.[30]

Antibodies specific to epigenetic marks can also be used for affinity enrichment of DNA regions. Antibodies can be used to immunoprecipitate modification-containing DNA fragments, allowing for their enrichment. This technique can then be coupled to methylation-specific oligonucleotide (MSO) microarrays, locus specific PCR, or whole genome approaches such as next generation sequencing or microarrays, to enable for low resolution mapping of epigenetic marks within the genome (see sections 1.4.4 and 1.4.5).

Figure 1.5. Scheme of ELISA methodology. In an ELISA the methylated DNA is coated onto the surface of a well-plate, incubated with a MeC specific antibody. Following binding of the primary antibody, a secondary antibody directed toward the primary antibody is added, which is conjugated to a dye or fluorophore. The amount of methylation can then be visualized by the level of absorbance or fluorescence intensity.



1.3.4 High performance liquid chromatography and mass spectrometry

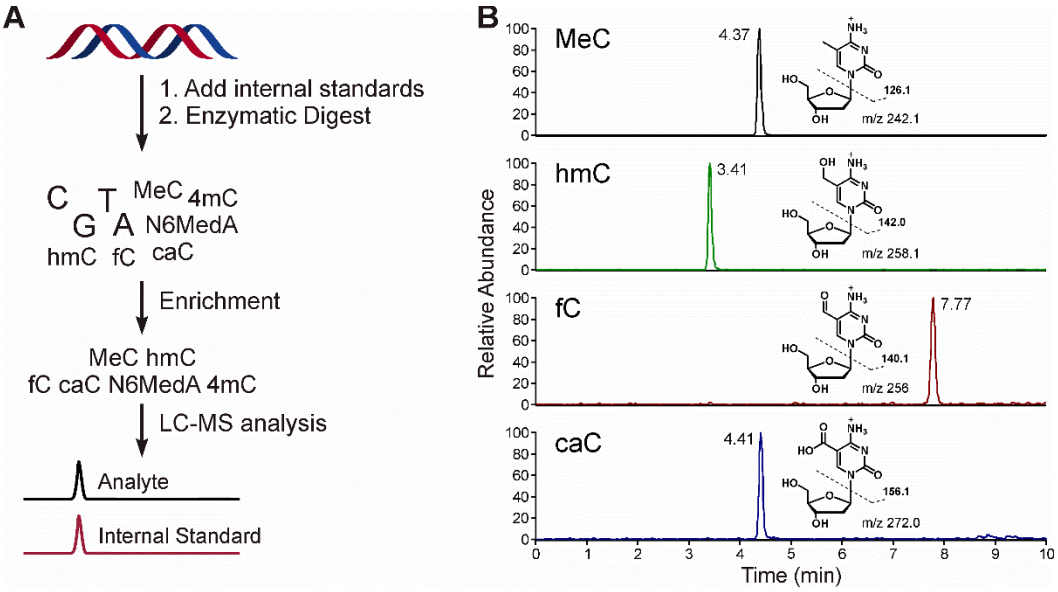
Global levels of MeC, hmC, fC, caC, N6MedA, and 4mC in genomic DNA can be accurately quantified by isotope dilution liquid chromatography mass spectrometry. In this approach, DNA is enzymatically digested using nucleases and phosphatases, and the resulting nucleosides are analyzed by HPLC-ESI-MS/MS. HPLC with UV detection can be used for measuring MeC in biological samples due to its high content (4-5% of all cytosine bases),[153] while other epigenetic modifications (hmC, fC, caC, N6MedA, and 4mC) are significantly less abundant and require more sensitive methods such as HPLC-ESI-MS/MS.[153]

HPLC-ESI-MS/MS is widely accepted as the most accurate technique for quantification of global levels of MeC and its oxidized forms.[153] In this approach, DNA is digested to individual nucleosides. Epigenetically modified DNA nucleosides are detected based on their characteristic molecular weights (**Table 1.1**), HPLC retention times, and MS/MS fragmentation (**Figure 1.6**). MeC, hmC, fC, caC, and N6MedA can be readily identified by mass spectrometry based on their distinct mass to charge ratios and MS/MS fragmentation.[153] Absolute quantitation is performed using stable isotope labeled analogues of each nucleoside that are spiked into samples (**Figure 1.6**), or by using external calibration curves. This methodology makes it possible to accurately quantify global amounts of each epigenetic mark of DNA, although it does not provide their distribution along the genome.[154]

Table 1.1. Molecular weights and main MS/MS fragments of canonical and epigenetically modified DNA nucleosides.

Nucleoside	**$[M + H]^+$**	**$[M + H - deoxyribose]^+$**
dC	228.0979	112.0505
dG	268.1040	152.0567
T	243.0975	127.0502
dA	252.1091	136.0618
MeC	242.1135	126.0662
hmC	258.1084	142.0611
fC	256.0928	140.0455
caC	272.0877	156.0404
N6MedA	266.1248	150.0774
4mC	242.1135	126.0662

Figure 1.6. Quantitation of epigenetic modifications via mass spectrometry. **A.** Mass spectrometry analyses begin with DNA that is spiked with internal standards and enzymatically digested to nucleosides. The hydrolysate is enriched for the target analytes prior to LC-MS analysis. **B.** Example of LC-MS/MS trace for the epigenetic modifications of cytosine, MeC, hmC, fC, and caC. Standards of MeC, hmC, fC, and caC, 100 fmol each were separated using a Luna Omega C18 column (Phenomenex, Torrance CA) which was eluted with a gradient of 0.1% acetic acid in water and pure acetonitrile. The mass spectrometric analysis was performed using a Thermo Fisher Scientific TSQ Quantiva mass spectrometer operating in SRM mode. The fragmentation conditions were optimized to give only the loss of deoxyribose as the primary fragmentation.



Methods that employ HPLC-ESI-MS/MS techniques to quantify epigenetic modifications have been developed for a variety of applications and instruments.[155-157] Due to the large number of methods that have been published for analyzing a combination of MeC, hmC, fC, and caC, only a few of them will be discussed here.

Earlier approaches employed offline analyte enrichment prior to HPLC-ESI-MS/MS analysis to minimize interference and improve detection sensitivity. The sample clean-up steps can include offline HPLC[15, 158] or solid phase extraction[20] In one example, DNA digests were subjected to offline HPLC clean-up, followed by HPLC-ESI-MS/MS/MS method on a LTQ linear ion trap.[158] The methodology of Liu et al. used 30-80 μg of genomic DNA and the limits of detection for hmC, fC, and caC were reported to be 2-3 modifications per 10^5 cytidines.[158]

More recent applications of HPLC-ESI-MS/MS for epigenetic marks of DNA have done away with offline sample clean-up steps or performed the enrichment online in a two-dimensional HPLC (2D-LC) tandem mass spectrometry method. A general method for analysis of MeC, hmC, fC, and caC was reported by Liu et al. which could be used at either analytical or nano HPLC flow rates.[153] DNA samples are enzymatically digested, followed by ultrafiltration to remove digestion enzymes. A column switching nano-HPLC methodology using two HPLC columns was used to allow for inline sample cleanup prior to high resolution MS/MS analysis using a Q-Exactive hybrid quadrupole-Orbitrap mass spectrometer.[153] Using this methodology, MeC, hmC, fC, and caC can be quantified at low femtomole levels.[153] A similar methodology was reported by Gackowski et al.[159] 2D-LC

allowed for isolation of peaks of interest (hmC, fC, and caC) as they eluted off the first HPLC column and directing them to a second column, which was connected to a Waters Xevo TQ-S mass spectrometer. Using this approach, a high level of sensitivity was achieved for hmC (LOQ: 3 fmol, LOQ: 10 fmol), fC (LOQ: 0.3 fmol, LOQ: 0.8 fmol), and caC (LOQ: 0.05 fmol, LOQ: 0.13 fmol) and the amounts of each were quantified in matched cancer and normal human colon tissue samples.[159] By minimizing sample processing and analyte clean-up steps, 2D-LC methodologies significantly simplify sample analyses, minimize the cost, allowing for high throughput quantification of large numbers of samples.

Although the vast majority of mass spectrometry based techniques do not allow for mapping epigenetic marks along the genome, a report by Lin et al. employed a MS approach to determine local concentrations of MeC at specific genes.[160] Following endonuclease cleavage with BbvI, DNA sequences of interest were captured by hybridization to complementary sequences immobilized on streptavidin beads.[160] The captured DNA was heated in formic acid to release MeC as a free base, which was detected using an LTQ mass spectrometer. Lin et al. used this method to probe methylation status of four promoter regions: GSTP1, BCL2, ESR1, and HIC1 in DNA from clinical prostate tissue samples digested with BbvI.[160] The capture methodology was able to effectively distinguish cancer tissue from normal tissue based on the level of promoter methylation of GSTP1, BCL2, ESR1, and HIC1.[160]

Biological samples from clinical studies are often precious, requiring method minimization to reduce DNA input. Le et al. have developed an HPLC-MS/MS method for the analyses of MeC and hmC in 50 ng of DNA.[161] This effort employed an Agilent 6460 QQQ, an Agilent Eclipse C18 (2.1 x 50 mm, 1.8 μm particle size) column eluted at a flow rate of 100 μL/min. External calibration curves were constructed using known concentrations of MeC and hmC. This method was able to quantify 0-10% MeC and 0-2% hmC.[161]

The recent discovery of N6MedA (**Figure 1.1**) in mammalian embryonic stem cells was accomplished by offline HPLC separation of enzymatic DNA digests, followed by HPLC-ESI-MS/MS analysis on a Waters UPLC coupled to a TSQ Quantum Ultra triple-quadrupole mass spectrometer.[36] Quantitation of the N6MedA was performed with external calibration curves, and the identity of the novel nucleoside in embryonic stem cells was confirmed using high resolution mass spectrometry.[36]

HPLC-ESI-MS/MS analysis of 4mC is not only challenging because it is present in relatively low amounts, but it also has the same molecular mass and similar biophysical properties as MeC (**Figure 1.1** and **Table 1.1**), making it difficult to detect it separately. HPLC resolution of 4mC and MeC was achieved by gradient elution of a C18 column under acidic conditions.[40] This allowed for 4mC quantification in the bacteria *C. kristjanssonii* which contains high concentrations of 4mC.[40]

1.4 Sequencing of epigenetic modifications in the genome

1.4.1 Indirect methods to map epigenetic DNA modifications along the genome

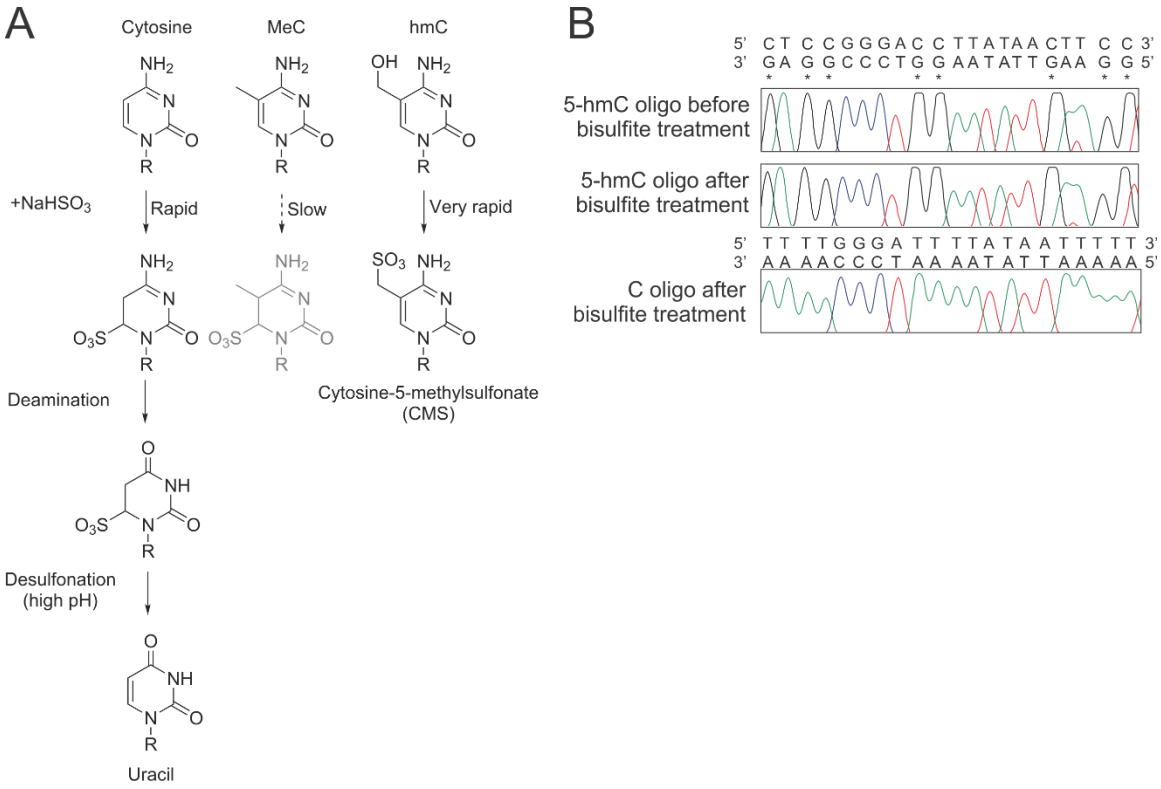
Mapping epigenetic DNA marks along the genome is essential for our understanding of their biological roles, interrelationships, and dynamics. The majority of techniques used currently employ bisulfite conversion, modified bisulfite conversion, or antibody capture to help locate epigenetic DNA modifications. Although the readouts from these techniques are indirect, they have yielded a wealth of important information over the years. Novel methodologies such as nanopore sequencing and single molecule real-time (SMRT) sequencing allow for a direct sequencing of modified DNA and are discussed in Section 1.5.

1.4.2 Bisulfite conversion to distinguish between C and MeC

Early studies of chemical reactivity of native and modified DNA nucleosides have revealed an important difference between C and MeC.[162-163] Upon treatment with bisulfite, cytosine is readily deaminated to uracil, while a much slower reaction is observed for MeC. The bisulfite reaction with cytosine proceeds through a C-6 sulfonate intermediate, which undergoes hydrolysis in the presence of base to release uracil (**Figure 1.7A**).[162] Unlike C, MeC is resistant to the deamination under the same conditions.[164] Therefore, bisulfite treatment of genomic DNA can be used to distinguish between methylated and unmethylated cytosines in the genome. Following bisulfite treatment, cytosine will be deaminated and subsequently read as T, while MeC will resist deamination and will be read

as C (**Figure 1.7A**). This unique property has been used by various sequencing technologies that employ bisulfite conversion to determine the locations of MeC bases in the genomes of various species including mammals, plants, and bacteria.[165-166]

Figure 1.7. Effects of bisulfite treatment on MeC and hmC in DNA and the sequencing results. **A.** Bisulfite-mediated deamination of cytosine. $HSO3^-$ adds across the 5,6-double bond of cytosine, promoting deamination and conversion to 6-sulfonyluracil. 6-sulfonyluracil is desulfonated to uracil (U) at higher pH. MeC is also deaminated to thymine by bisulfite conversion, but the rate is approximately two orders of magnitude slower than that of cytosine. Bisulfite quickly converts 5-hydroxymethylcytosine to form cytosine-5-methylenesulfonate (CMS), which does not readily undergo deamination. **B.** Shown are examples of Sanger sequencing traces of hmC-containing DNA before and after bisulfite treatment. hmC resists bisulfite induced-deamination and is thus read as cytosine. The control C-containing oligonucleotide, on the other hand, shows complete conversion of all C's in the top strand (highlighted sequences) to T's (lower panel). Adapted from Huang Y et al. (2010) The behavior of 5-hydroxymethyl-cytosine in bisulfite sequencing. PLoS One. Jan 26;5(1):e8888. doi: 10.1371/journal.pone.0008888. Used in accordance with CC BY license (http://creativecommons.org/licenses/by/4.0/).[167]


A
Cytosine
MeC
hmC
+NaHSO3
Rapid
Slow
Very rapid
Deamination
Cytosine-5-methylsulfonate
(CMS)
Desulfonation
(high pH)
Uracil
B
5' C T C C G G G A C C T T A T A A C T T C C 3'
3' G A G G C C C T G G A A T A T T G A A G G 5'
5-hmC oligo before
bisulfite treatment
5-hmC oligo after
bisulfite treatment
5' T T T T G G G A T T T T A T A A T T T T T T 3'
3' A A A A C C C T A A A A T A T T A A A A A A 5'
C oligo after
bisulfite treatment

1.4.3 Microarray technology for sequencing

A DNA microarray is a collection of DNA segments of known sequence attached to a solid surface in a known pattern. DNA to be sequenced is fragmented and modified with fluorescent tags incorporated through PCR. DNA fragments present in DNA samples are hybridized to specific sites on the array and can be fluorescently imaged.[168] The extent of enrichment can be quantified from the intensity of the fluorescent signal.[168] An upstream enrichment step is a necessary component in using DNA micro arrays. As such, micro arrays are often coupled to antibody enrichment, or used with cDNA to explore RNA abundance under different treatment conditions. The microarray methodologies require small sample input and are easily replicated for multiple samples.

In 2002, Gitan et al. developed the first methylation-specific oligonucleotide (MSO) microarray for DNA methylation analysis.[169] In this technique, DNA samples are treated with bisulfite to convert cytosine to uracil, whereas MeC remains unchanged (**Figure 1.7A**). Specific CpG containing regions of the genome are then amplified by PCR and labeled using fluorescent dye-conjugated primers and hybridized to oligonucleotide probes attached to a glass surface. The oligonucleotide probes are specifically designed to discriminate between bisulfite converted and unconverted sequences containing known CpG sites. The quantitative differences in hybridization, which can be assessed by the fluorescent intensity, indicate the methylation status of a particular genomic locus. This first use of a MSO microarray successfully mapped methylated CpG sites within the estrogen receptor gene CpG island in cultured breast cancer cells and clinical samples.[169]

Further developments in the MSO microarray methodologies have led to a creation of gene-specific, multiplex systems that can be used to simultaneously examine methylation across multiple CpG sites of numerous genes. This method provides both qualitative and quantitative DNA methylation data.[170] Due to these advantages, the MSO microarray technology was commercialized and is now widely used to study cytosine methylation along the genome.

A variety of commercially available methylation microarrays are available from companies such as Illumina, Affymetrix, Agilent, and Roche NimbleGen. One such commercially available microarray is the Infinium Human Methylation 450 BeadChip (Infinium Methylation 450K) by Illumina. The Infinium 450K makes it possible to assess the methylation status of more than 480,000 cytosines distributed over the entire human genome.[171] This microarray utilizes a similar technique as the MSO microarray developed by Gitan et al. in 2002 and can determine the methylation status of 96% of CpG islands and 99% of known human genes using ~500 ng DNA.[148] The CpG sites being interrogated include promoter regions, 5′-UTRs, gene bodies, 3′-UTRs, and intergenic regions.[171] The Infinium 450K is also applicable to both cell and tissue samples, however this array lacks coverage of distal regulatory elements.[172]

Although the Infinium 450K had been widely used by the epigenetics research community, the technology has presented some technical challenges.[172] It was reported that many of the probes in the Illumina 450k array either cross-hybridize to non-targeted genomic regions or target loci that contain known single nucleotide polymorphisms

(SNPs), both of which interfere with analysis of DNA methylation levels.[173] Because of this, sites of differential DNA methylation identified using the Illumina 450k array must be confirmed using a second independent assay, especially the sites targeted by the cross-reactive probes or that contain known SNPs. This led Illumina to release a new technology, the Infinium MethylationEpic (EPIC) BeadChip. This new array was specifically designed to target enhancer regions. It contains over 850,000 probes, which cover more than 90% of the sites on the Infinium 450K, plus more than 350,000 CpGs at regions identified as potential enhancers by FANTOM5 and the ENCODE project.[174-175] In this array, probe sequences are designed to be complementary to specific 50 base regions of bisulfite converted genomic DNA with a CpG site at the 3′ end of the probe. Following hybridization to bisulfite converted DNA, single-base extension of the probe incorporates a fluorescently labelled ddNTP at the 3′ CpG site to distinguish between the C/T conversion that results from bisulfite treatment.[172]

Methylation array methodologies described above represent a cost-effective and high-throughput technology to quantify the methylation level of half a million cytosines all over the genome. Furthermore, these arrays have low input DNA requirements, making them a powerful tool to study DNA methylomes. However, there are potential biases with this technology. As with all bisulfite-based methods, incomplete bisulfite conversion can affect quantitation of methylation differences. In addition, single nucleotide polymorphisms (SNPs) can interfere with methylation level detection, and commercial arrays are limited to human samples only. No commercial arrays are now available to map

methylation of the murine genome, hindering the application of this methodology to the wide range of valuable animal models of disease.[176-177] Despite these limitations, Needhamsen et al. recently reported the use of the Infinium Human MethylationEPIC BeadChip to study DNA methylation in mouse DNA samples.[177] This study identified over 19,000 EPIC probes which aligned to the bisulfite converted reference mouse genome.[177] Another limitation of this technology, is its inability to differentiate methylation from hydroxymethylation.[167]

Recent work by Nazor et al, combines TET-assisted bisulfite (TAB) conversion with Illumina 450K DNA methylation arrays. This method, termed TAB-array, provides a low-cost, high-throughput approach that distinguishes hmC from MeC signals at base resolution.[178] This technique requires that the genomic DNA from each biological sample is split in two fractions for different downstream processing steps. One fraction undergoes normal bisulfite conversion which identifies cytosines as methylated or not methylated, without distinguishing between MeC and hmC. The other fraction undergoes TAB conversion in which DNA is first glucosylated to protect hmC, and then oxidized by TET1, followed by bisulfite treatment. This allows hmC to be distinguished from MeC and quantified at the CpG sites included in the 450K array.[178]

1.4.4 Next generation sequencing technology

Next generation sequencing includes techniques that utilize polymerase synthesis of complimentary DNA strand to determine the sequence of input DNA. The technology

used by Illumina (San Diego, CA) immobilizes DNA fragments onto a chip by hybridizing them to synthetic DNA oligomers that were seeded on the chip during manufacturing. The immobilized DNA is then amplified on the chip to create clusters of identical DNA fragments, which will increase the recorded signals from DNA synthesis. Following amplification, the DNA clusters are sequenced by synthesis of the complementary DNA using engineered DNA polymerases and specially modified nucleic acids.[179] Each nucleotide has a fluorescent tag and is reversibly blocked modified at the 3′ end to allow for only a single incorporation.[179] Following single nucleotide addition, a camera images the chip to record a fluorescent signal and records the sequence being made, followed by deprotection of the 3′ end for continued synthesis.[179] These steps for elongation are repeated until the desired read length is achieved. This technology can be combined with antibody/enrichment-based sequencing (Section 1.4.5), whole genome sequencing (Section 1.4.6), reduced representation sequencing (Section 1.4.7), and modified bisulfite sequencing (Section 1.4.8) as detailed below.

1.4.5 Antibody/Enrichment-based sequencing

Low resolution data for genome wide distribution of DNA methylation marks can be obtained using antibody-based techniques such as Methylated DNA Immunoprecipitation (MeDIP). In this assay, MeC targeted antibody is used to specifically immunocapture methylated genomic fragments.[180] The enriched DNA fragments can then be input into high throughput DNA detection methods such as DNA microarray (MeDIP-

ChIP) or next generation sequencing (MeDIP-seq). Through genome-wide peak calling of the microarray or the respective sequencing results, methylated regions along the whole genome can be identified.[180-181] The advantages of this method are its low cost and the ability to rapidly identify DNA methylation regions within the genome.

Similar immunoprecipitation (IP) based methodologies have been developed for oxidized forms of MeC (including hmC, fC, and caC), since the corresponding antibodies against them are now readily available.[42, 182] This approach has been used in multiple studies to identify regions of hmC enrichment such as CpG rich transcription start sites and fC and caC enrichment at pericentric heterochromatin, as well as at distal regulatory elements.[42, 182]

Apart from direct antibody enrichment techniques, a number of chemically assisted techniques have been developed for mapping hmC and fC. In anti-cytosine-5-methylenesulfonate (anti-CMS Seq) technique, hmC is converted to cytosine-5-methylenesulfonate (CMS) by bisulfite treatment, followed by enrichment of CMS-containing DNA fragments via CMS-targeted antibody and their sequencing.[183] This methodology has a lower background and lower reliance on hmC density in the fragment as compared to direct antibody pull-down of hmC containing fragments.[183-184]

In GLIB-seq (Glucosylation, periodate oxidation, biotinylation Sequencing), hmC is glucosylated in the presence of β-glucosyltransferase, and the resulting glucosyl derivatives are oxidized with sodium periodate to yield two aldehyde groups.[183] The latter are derivatized with biotin-containing aldehyde reactive probe (ARP), and the derivatized

fragments are enriched via streptavidin beads and sequenced, allowing for mapping of hmC marks along the genome.[183] Another chemically assisted enrichment method for hmC is hMeSeal-seq (hydroxymethyl selective chemical labeling).[185] In this method, an engineered glucose moiety containing an azide group is enzymatically added to the hydroxyl group of hmC by T4 bacteriophage β-glucosyltransferase. The azide group can then be chemically modified with biotin for detection, affinity enrichment, and sequencing hmC–containing DNA fragments.[186]

Similar mapping methodologies have been developed for fC.[43, 187] Since fC contains an inherently reactive aldehyde group, it can be directly modified with biotin-containing aldehyde reactive probe, followed by enrichment of fC containing DNA fragments using streptavidin beads.[43] Another method for mapping fC involves enzymatically blocking hmC with unmodified UDP-Glc.[187] This is followed by fC reduction to hmC with $NaBH_4$ and hMeSeal-seq to localize the newly formed hmC sites as fC fragments.[187]

The main limitation of all enrichment-based methods is their low resolution (100-300 bp).[188] This can be a major problem, especially for studies utilizing multiple epigenetic marks. Another limitation is that these methods are unable to quantify absolute methylation levels because each enriched fragment may have variable numbers of epigenetic marks. Additional common concerns include antibody cross-reactivity, low capture efficiency, and the limitations of downstream DNA detection methods (microarrays and sequencing) used to determine the enrichment. However, these techniques are still widely used as they

provide a rapid, comparatively less laborious and inexpensive way to study genome wide methylation.

1.4.6 Whole genome bisulfite sequencing

Studies of DNA methylation were initially restricted to localized CpG-rich regions of the genome.[189] However, several methods now exist which can map DNA methylation genome wide. Whole-genome bisulfite sequencing (WGBS) was developed in 1992 by Frommer et al. to measure CpG methylation genome-wide at a single-base resolution.[190] This method utilizes bisulfite-induced modification of genomic DNA to convert cytosine to uracil, the latter is converted to thymine during PCR amplification. MeC residues however, are not converted and remain as cytosines after PCR amplification.[190] The development of high-throughput sequencing has facilitated the generation of genome-wide, single-base resolution DNA methylation maps from bisulfite-converted DNA.[191] The first methylome was generated in 2008 from the *Arabidopsis thaliana* genome.[192-193] The first human methylome was reported by Lister et al. in 2009 from human embryonic stem cells and fetal fibroblasts.[194] These studies revealed the genome wide context of DNA methylation. Extensive differences were found between the methylomes of two human cell types, revealing the dynamic nature of this modification. Due to the high genomic coverage of WGBS (95%), this method is preferred for building reference methylomes.[195]

1.4.7 Reduced representation bisulfite sequencing (RRBS)

Although WGBS allows for the greatest coverage of DNA methylation genome-wide, it is time consuming and requires significant financial investment in sequencing to reach the effective depth required for accurate DNA methylation mappings; this hinders its application to large-scale DNA methylation studies.[196] Reduced representation bisulfite sequencing (RRBS) is an efficient and high-throughput technique used to analyze methylation profiles at a single-nucleotide level from regions of high CpG content.[197-198] This technique utilizes a methylation-insensitive restriction enzyme, such as MspI, which cuts at 5′-CCGG-3′ sites, enriching for CpG rich regions of the genome such as CpG islands by leaving CpG rich regions at the ends of fragments being sequenced.[191] This portion of the genome is then sequenced using NGS to generate single-base pair resolution DNA methylation maps. RRBS captures 85% of CpG islands and 60% of gene promoters, requiring small sample input (0.01-0.03 μg of DNA).[191, 199] By combining different restriction enzymes, CpG coverage across the genome can be altered to include or exclude certain regions of interest such as CpG island shores, which are the two kilobase regions adjacent to CpG islands and are known to play an important role in various biological processes including cellular differentiation.[200-201]

One of the main advantages of RRBS over WGBS is the overall reduction of the size of DNA regions to be sequenced, therefore requiring fewer reads for accurate sequencing reducing the cost and allowing for deep sequencing.[191] An important disadvantage of RRBS is that it does not capture all CpG islands and/or promoters. While

WGBS can access ~95% of the genome, RRBS only reaches 3.7% genome coverage.[191] However, RRBS enjoys much better coverage depth than WGBS and the concordance of DNA methylation levels between the two methods is high.[196] Due to these advantages, RRBS has become popular when high-throughput, low cost methylation analysis is needed such as in clinical applications.[202] It is important to note however, that RRBS has been mostly applied for DNA methylation profiling of CpG sites within mouse and human genomes, and the ability of MspI coupled RRBS to cover non-CpG methylation has not been demonstrated.[196]

1.4.8 Modified bisulfite sequencing to facilitate detection of hmC, fC and caC by next generation sequencing

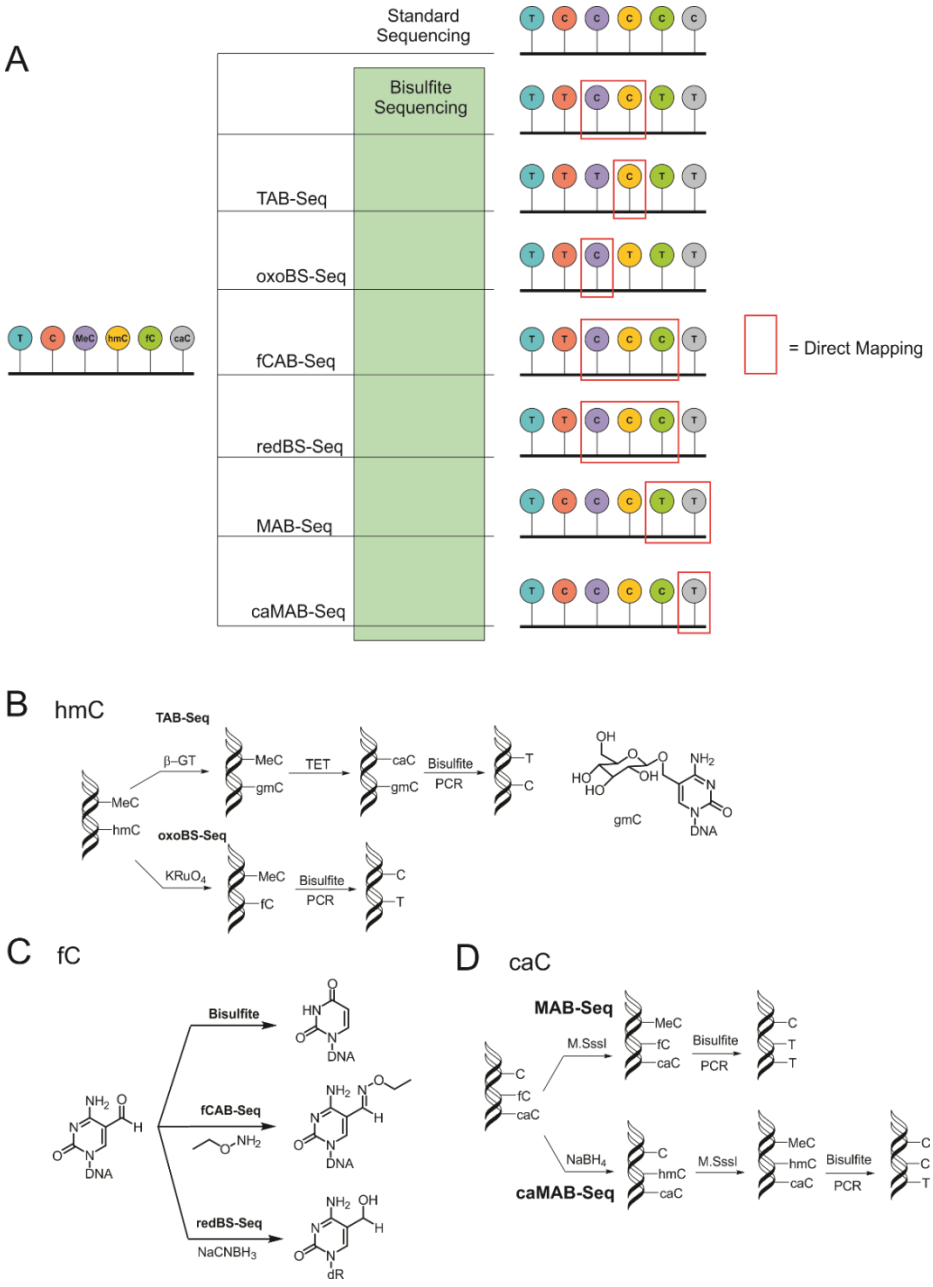
The discovery of hmC, fC, and caC as oxidation products of MeC generated by TET enzymes has spurred a flurry of speculation about possible biological roles of these additional epigenetic marks.[203] In addition to their role as demethylation intermediates, these oxidized forms of MeC may serve as unique epigenetic marks in their own right, interacting with a distinct group of protein readers and eliciting a unique epigenetic response.[26-27] Therefore, it is of great interest to map the locations of hmC, fC, and caC along the genome. Similarly to MeC, hmC resists bisulfite mediated deamination, thus the two marks cannot be distinguished by standard bisulfite methodology (**Figure 1.7**).[164, 167] [167, 204] Therefore, previous studies that employed bisulfite conversion to map MeC actually detected both hydroxymethylation and methylation together. In contrast, fC and caC

readily undergo deamination upon treatment with bisulfite and thus cannot be distinguished from C (**Figure 1.8**).[15-16, 205] In addition, fC also undergoes deformylation which is mediated by sulfonation of the C6 position of the nucleoside to generate the corresponding alcohol (**Figure 1.8**). Base catalyzed deprotonation of the alcohol and β-elimination produce cytosine.[16] Determining the levels and positions of epigenetic modifications in the genome is important to understanding the biological function of these DNA marks. Therefore, novel sequencing methodologies were needed to separately map all four epigenetic marks.

The first method for distinguishing between MeC and hmC during bisulfite sequencing utilized an additional chemical reaction on the genomic DNA. The oxidative bisulfite sequencing (oxBS-Seq) includes DNA treatment with potassium perruthenate ($KRuO_4$) to selectively oxidize the hmC to fC (**Figure 1.8**).[205] After the oxidation, the DNA is treated with bisulfite to deaminate cytosine. Booth et al. coupled oxBS-Seq to RRBS to increase the sequencing depth at potentially interesting CG repeats by reducing the amount of DNA that could be sequenced. hmC was identified primarily at CG sites located within low density CpG islands (CGIs).[205] In addition to chemical oxidation to distinguish between MeC and hmC enzymatic oxidation can used (TET Assisted Bisulfite Sequencing, TAB-Seq).[206] The latter technique utilizes T4 β-glycosyl transferase and UDP-glucose to protect the primary alcohol on hmC, followed by oxidation of MeC to caC using recombinant TET1 protein (**Figure 1.8**).[207] This is coupled with bisulfite treatment and sequencing of the original sample and Tet treated DNA to provide data for MeC and hmC

location. Using the technique on the whole genome, hmC was found to be almost exclusively found at CG sites, and nearly half of all hmC sites were present in distal regulatory elements.[206] TAB-Seq has been applied to identify changes in MeC and hmC in intergenic regions of kidney tumors[208] and to identify sites of 4mC in bacterial DNA.[40]

Figure 1.8. Methodologies for distinguishing epigenetic modifications in DNA using bisulfite treated DNA. **A**. Diagram of the Bisulfite coupled sequencing technologies used for mapping the positions of MeC, hmC, fC, and caC. **B**. Sample preparation scheme for differentiating MeC and hmC in TAB-Seq[206] and oxBS-Seq.[205] **C**. fC is deaminated and deformylated by treatment with bisulfite. Protection of fC by reaction with O-ethylhydroxylamine in fCAB-Seq prevents the deamination and deformylation that typically occur when fC is treated with bisulfite.[187] In redBS-Seq, fC is reduced with sodium cyanoborohydride to hmC which will resist deamination.[209] fC in the original sample can be determined by comparison of fCAB-Seq or redBS-Seq data with DNA that had undergone standard bisulfite sequencing. **D**. Because fC and caC deaminate when treated with bisulfite, MAB-Seq utilizes the methylase M.SssI to methylate unmodified cytosines. Following bisulfite treatment, MeC and hmC will be identified as C, while fC and caC are deaminated and identified as T their levels and position in the genome.[210] caMAB-Seq first reduces fC to hmC and then follows MAB-Seq in order to specifically map caC.[210]


A
Standard Sequencing
Bisulfite Sequencing
TAB-Seq
oxoBS-Seq
fCAB-Seq
redBS-Seq
MAB-Seq
caMAB-Seq
T C MeC hmC fC caC
= Direct Mapping
B hmC
TAB-Seq
β-GT
MeC
hmC
TET
gmC
caC
Bisulfite PCR
oxoBS-Seq
KRuO4
fC
gmC
DNA
C fC
Bisulfite
fCAB-Seq
redBS-Seq
NaCNBH3
dR
D caC
MAB-Seq
M.SssI
caC
NaBH4
caMAB-Seq

Mapping fC along the genome was accomplished in part by modification of the previous techniques to distinguish between MeC and hmC. The first reported method for fC was formylcytosine chemical assisted bisulfite sequencing (fCAB-Seq).[187] In this approach, chemical labeling of fC is conducted with *O*-ethylhydroxylamine to protect fC from conversion during DNA treatment with bisulfite (**Figure 1.8**).[187, 205] DNA sequencing is performed on paired samples that had undergone fCAB protection plus bisulfite treatment and standard bisulfite treatment. Through the mapping of standard bisulfite sequencing results and fCAB-Seq data to the genome, sites specifically modified with fC could be identified. Cytosine formylation amounts at these sites can be estimated by subtracting the percent cytosine methylation determined in the bisulfite only treated sample (which is the sum of MeC and hmC) from the percent of cytosine methylation in the fCAB-Seq sample data (which is the sum of MeC, hmC, and fC).[187] Because fC is a minor component of DNA, each region of DNA was sequenced an average of 1000 times, which allows estimation of fC levels to +/- 0.1%. Sequencing of DNA from mouse embryonic stem cells showed that fC levels were increased at poised enhancers sequences, while MeC was also reduced at those same enhancers, indicating MeC oxidation by TET could function in epigenetic priming.[187] Another chemical method for identification of fC is reduced bisulfite sequencing (redBS-Seq). The redBS-Seq method employs sodium cyanoborohydride to reduce fC to hmC in genomic DNA (**Figure 1.8**).[209] The DNA is then treated with bisulfite to deaminate cytosine and both redBS and standard bisulfite treated DNA samples are sequenced. Cytosine formylation levels are determined as in fCAB-Seq

above. Using this technique, fC was shown to be present in CpG sites asymmetrically with hmC, where one strand contains fC and the opposite strand hmC.[209]

An additional method for determination of oxidized forms of MeC by sequencing is modification-assisted bisulfite sequencing (MAB-Seq) (**Figure 1.8**).[210] Detection of caC and fC is simultaneously possible by treatment of the DNA with bacterial methylase *M.SssI*, which methylates all unmodified CG dinucleotides to give MeCG. Following methylation to protect all unmodified cytosine in a CG context, bisulfite treatment of DNA will lead to deamination of fC and caC. This results in sequencing output that can identify the locations of fC and caC together. MAB-Seq is reliant on quantitative conversion of C to MeC by *M.SssI* and is limited to detection of fC and caC only in the CG context. MAB-Seq was further developed to be specific for caC (caMAB-Seq) (**Figure 1.8**), which takes advantage of the reduction of fC by sodium borohydride prior to performing the steps of MAB-Seq.[210] Using MAB-Seq and caMAB-Seq, fC and caC have been mapped at single base resolution to regions with higher chromatin accessibility, and therefore TET accessibility, despite TET stalling on hmC during oxidation.[18] The sequencing profiles obtained via MAB-Seq and caMAB-Seq indicate that there might be preferential locations for TET activity, either through chromatin accessibility or sequence dependent removal of fC and caC by TDG.[210]

The first sequencing of 4mC in the genome was reported using a modification of the TAB-Seq methodology.[40] This is possible because 4mC undergoes deamination at a rate between C and MeC.[40] To allow for 4mC detection, the bisulfite conditions were

optimized to ensure the retention of 4mC in the genome while deaminating TET-generated caC. Using this methodology, 4mC was mapped in *C. kristjanssonii*, a bacterium with similar levels of MeC and 4mC.[40] This methodology, 4mC-TAB-Seq, is well suited for identification of 4mC in genomes which have high concentrations of 4mC at particular sites, especially in bacteria which utilize 4mC as part of their restriction-modification systems.[38] Application of this methodology to mammalian cells and tissues is limited by the incomplete retention of 4mC after bisulfite conversion (40-55%), making accurate identification of sites possessing less than complete methylation nearly impossible.[40]

Mapping N6MedA can also be accomplished using restriction enzyme-based sequencing (6mA-RE-Seq).[211] This sequencing technique takes advantage of previous information gained from antibody enrichment methods, which found that N6MedA was found in CATG and CATC motifs in the *C. reinhardtii* genome, a green algae.[211] Focusing on these two motifs, 6mA-RE-Seq employs digestion of DNA by CviAII or DpnII, sonication to fragment the DNA, end repair, adenylation, and ligated to adapters.[211] The methylated sequences will resist enzymatic digestion, allowing indirect detection of N6MedA based on the location of the motif in the sequencing read. The motif at the end of a read indicates a lack of methylation, while a read with the motif in the middle indicates the presence of N6MedA. The amount of N6MedA can be calculated from the relative number of sequencing reads where the specific adenosine is located at the end of a read (unmethylated) or at an internal position (methylated). Using this technique, N6MedA was identified at 3-4% of all available motifs in the *C. reinhardtii* genome.[211] When N6MedA

containing sites identified by 6mA-RE-Seq were compared to those from immunoprecipitation (IP), 28% of sites identified by IP were not identified by 6mA-RE-Seq and did not contain the motifs used in 6mA-RE-Seq. Using 6mA-RE-Seq and IP sequencing Fu et al. found that N6MedA was enriched in the regions of DNA linking nucleosomes possibly used for positioning these nucleosomes on DNA.[211]

1.5 Direct detection and mapping of epigenetic modifications in DNA

Unlike methodologies described above and summarized in **Table 1.2**, third generation sequencing methods allow for reading the nucleotide sequences at the single molecule level without additional modification reactions like bisulfite conversion, making it possible to directly distinguish between C, MeC, and its oxidized forms during the sequencing step. These methodologies are currently under active development.

Table 1.2. Brief description of techniques for performing base resolution sequencing of MeC, its oxidized forms, 4mC, and N6MedA.

Sequencing Technique	Description
WGBS	Bisulfite treatment and sequencing are carried out on the whole genome. Reads both MeC and hmC as C after bisulfite conversion.
RRBS	CpGs are recognized and cut by specific enzymes. Traditional bisulfite sequencing is carried out on this enriched CpGs region. Reads both MeC and hmC as C after bisulfite conversion.
TAB-Seq	hmC is converted to 5gmC by the β-GT enzyme. This protects hmC from TET oxidation, which is used to convert MeC to caC. hmC is detected after bisulfite sequencing. MeC is determined by subtracting hmC from standard BS-Seq.
oxBS-Seq	hmC is treated with potassium perruthenate and converted to fC. This will cause fC to be read as T after bisulfite conversion. hmC identified by subtracting the signal for MeC from traditional BS-Seq.
redBS-Seq	fC is reduced to hmC with sodium borohydride. This will cause fC to be read as C in BS-Seq. The sites of fC can be determined by subtracting BS-Seq methylation from redBS-Seq.
fCAB-Seq	fC is reacted with $EtONH_2$, which protects fC from bisulfite mediated deamination. This allows fC to be read as C in BS-Seq. The location of fC can be identified by subtracting BS-Seq methylation from fCAB-Seq.
MAB-Seq	Unmodified CpGs are methylated by M.SssI. fC and caC can then be identified in BS-Seq by C to T transition in the sequencing data relative to the reference genome.
caMAB-Seq	fC is reduced to hmC with sodium borohydride. Unmodified CpGs are methylated by M.SssI. caC can then be identified in BS-Seq by C to T transition in the sequence relative to the reference genome.

1.5.1 Single molecule real-time (SMRT) sequencing

As DNA sequencing methodologies entered the third generation, single molecule real-time (SMRT) technology has become a new favorite since it has longer read lengths (10,000 bp up to 60,000 bp) produces higher consensus accuracy, and is characterized by a lower degree of bias as compared to second generation sequencing.[212] Unlike the aforementioned sequencing methods, SMRT-seq can directly read different base modifications via the kinetics of nucleotide incorporation during the normal course of sequencing of an intact DNA.[213] In SMRT-seq, kinetics signatures of MeC and hmC are very subtle, presenting an accuracy challenge for detection, whereas fC and caC show very distinct kinetics signatures.[214] More recently, it was shown that glycosylation of hmC significantly enhances its kinetics signature, making simultaneous mapping of all three oxidized forms of MeC possible.[215]

1.5.2 Nanopore sequencing

Another emerging third generation sequencing technology is Nanopore sequencing, which can directly read out different bases using distinct current signals generated as DNA translocates through a protein nanopore.[216] Generally, a protein nanopore is set in an electrically resistant polymer membrane and an ionic current is passed through the nanopore by setting a voltage across this membrane. When DNA passes through the pore or near its aperture, this event creates a characteristic disruption in the current. Measurement of that current makes it possible to identify the molecule in question.

Previous study has shown the ability of discriminating between all 5 cytosine variants by a low-throughput nanopore sensors,[217] although high throughput nanopore sequencing can only discriminate cytosine, MeC and hmC.[218]

1.6 Challenges of using sequencing to identify epigenetic modifications at single base resolution

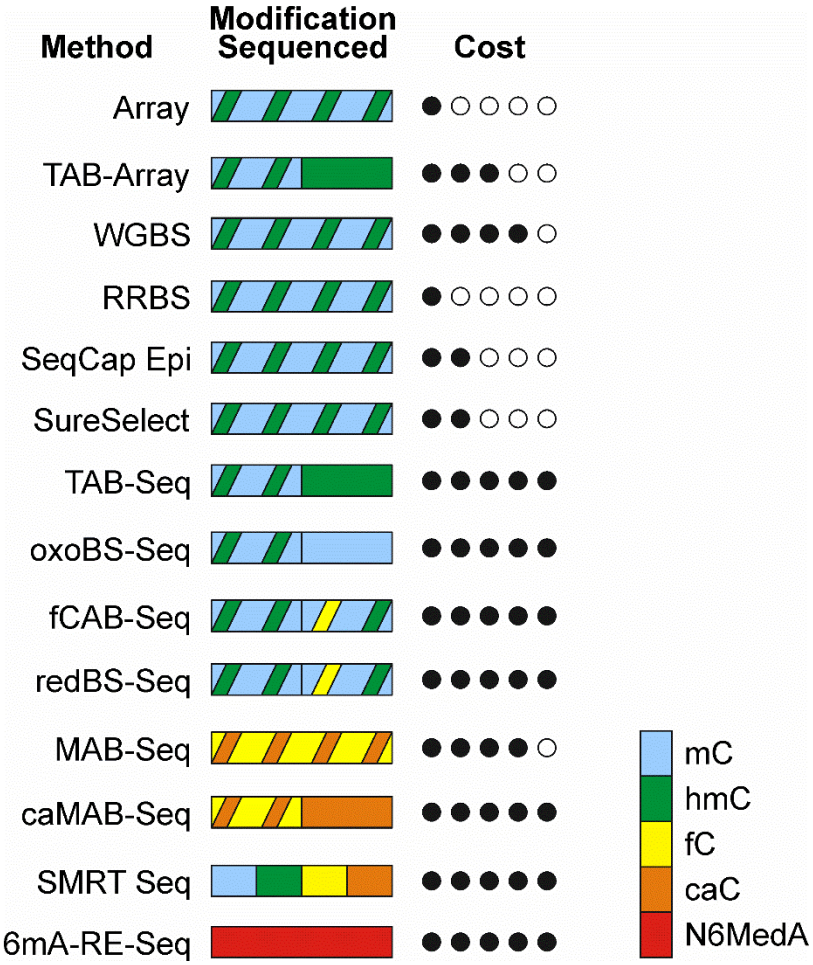
As described above, application of DNA sequencing techniques can provide base resolution maps of DNA epigenetic marks, allowing for elucidation of their biological functions.[205-206, 210] However, significant challenges remain in applying sequencing strategies to map non-canonical DNA bases. Sequencing cost is the first factor that needs to be considered when planning a new study. Although the cost of whole genome sequencing has decreased significantly in recent years, it remains around $1000 per genome according to the National Human Genome Research Institute,[219] making this approach non-practical for large studies involving hundreds of samples (**Figure 1.9**).[220] One possible solution to the high cost of whole genome sequencing includes RRBS, however RRBS introduces an additional challenge with a more limited coverage of CpG regions.[196] MspI is a restriction enzyme commonly used in RRBS methods, however theoretically it can only cover 10% of the genomic regions.[199] More targeted capture based approaches are available for sequencing specific regions, such as the SeqCap Epi (Roche) and SureSelect (Agilent) systems.[221-222] These capture methods focus on predetermined genomic regions in order to reduce sequencing complexity, but are by design limited

around specific genomic regions which may or may not include the regions of interest. Custom capture libraries can be created if certain regions are not part of the original capture systems.[223]

Secondly, for the methods relying on bisulfite sequencing, the conversion rate of the target bases remains to be improved. Without excellent, specific conversion the uncertainty in mapping target bases will significantly affect the estimation of their abundance in particular regions. In addition, these methods usually involve several rounds of sequencing under different conditions, which increases the cost, labor, as well as requiring increased DNA sample input. Due to the deamination of cytosine to uracil during the sample preparation, these methods also pose additional challenges after sequencing including read alignment and bioinformatics analysis.

Lastly, third generation sequencing technologies such as single molecule real-time (SMRT) sequencing and Nanopore sequencing permit longer reads and direct detection of epigenetic modifications of DNA. Because these methods do not rely on DNA conversion with bisulfite or enzymatic processes and allow for longer reads, read alignment is simplified. Despite these positives, third generation sequencing still needs to improve, especially in regard to the signal intensity for some specific epigenetic marks, the overall accuracy, and larger DNA input requirements. These limitations prevent the application of this technology to whole genome sequencing without some form of enrichment.[36]

Figure 1.9. The effectiveness of various sequencing technologies at discriminating between epigenetic modifications in DNA and the relative cost of their application. The ability of a method to identify MeC (Blue), hmC, (green, both MeC and hmC (blue and green stripes), fC (yellow), caC (orange), both fC and caC (yellow and orange stripes), and N6medA (red). The relative cost is depicted as filled circles, least expensive (●○○○○) to most expensive (●●●●●).



1.7 Summary and thesis goals

Epigenetic regulation of gene expression is central to controlling physiological processes in cells and is mediated through covalent DNA, histone, and RNA modifications and non-coding RNAs.[1] The DNA epigenetic modifications include MeC, hmC, fC, caC, and N6MedA (**Figure 1.1**). Epigenetic DNA modifications are placed at specific sites enzymatically. These epigenetic modifications affect the transcription of genes and need to be faithfully maintained through replication cycles.[1] Patterns of epigenetic modifications, which are responsible for stable gene expression and cell identity can be altered by cigarette smoke exposure and by inflammation.[141, 224] These epigenetic DNA modifications can be quantified post-exposure to measure the magnitude of change in the global epigenetic levels, providing an overview of how the whole genome was affected. The oxidation of MeC to hmC, fC, and caC by TET proteins provided potential new epigenetic marks; an identity that has been supported by additional reports.[26-27] However, the kinetic effects of these epigenetic DNA modifications on DNMT1 activity had not been explored. Thus, the initial goal of this thesis was to develop a kinetic assay of DNMT1 to investigate its activity when presented with alternative substrates.

We developed an *in vitro*, kinetic assay using isotope dilution tandem mass spectrometry for the quantitation of MeC formation in digests of synthetic DNA duplexes incubated with DNMT1. Passive demethylation, where DNMT1 fails to properly methylate the newly synthesized DNA has been suggested as a mechanism to facilitate DNA

demethylation. We tested whether DNMT1 would effectively recognize the oxidized forms and methylate the duplex accordingly.

Additional studies were focused on animal models of smoking induced lung cancer. We developed an isotope dilution, capillary-LC tandem mass spectrometry method to quantify global epigenetic changes in mice treated with NNK, LPS, and both NNK and LPS. The exposure of A/J mice to cigarette smoke is hypothesized to capture the tumor initiation pathway of smoking in humans. The type II alveolar cells are hypothesized as sites of lung cancer initiation due to their proliferative ability.[225] We tested this hypothesis by determining the gene expression profile of isolated type II cells from A/J mice and performing oxoRRBS to site specifically assign changes in DNA methylation and hydroxymethylation in mice exposed to cigarette smoke.

To further our understanding of the processes and proteins associated with the oxidized forms of MeC in DNA we began development of a protein pull-down assay to identify epigenetic readers in the lung.

II. HUMAN DNA METHYLTRANSFERASE ACTIVITY IN THE PRESENCE OF 5-METHYLCYTOSINE ANALOGUES WITH EXTENDED ALKYL SIDE CHAIN

Adapted from:

Kotandeniya, D.; Seiler, C. L.; Fernandez, J.; Pujari, S. S.; Curwick, L.; Murphy, K.; Wickramaratne, S.; Yan, S.; Murphy, D.; Sham, Yuk Y.; and Tretyakova, N. Y.; Can 5-methylcytosine analogues with extended alkyl side chain guide DNA methylation? *Chem. Commun.* **2018**, 54, 1061-1064.

This work was performed in collaboration with Dr. Delshanee Kotandeniya, Jenna Fernandez, Dr. Suresh S. Pujari, Lauren Curwick, Kristopher Murphy, Dr. Susith Wickramaratne, Shuo Yan, Dan Murphy, and Dr. Yuk Y. Sham, under the direction of Dr. Natalia Tretyakova. Delshanee Kotandeniya and Christopher Seiler performed the kinetics experiments for cytosine, ethyl, propyl, and methylcytosine containing DNA. Christopher Seiler performed the kinetics experiments for vinylcytosine containing DNA. Delshanee Kotandeniya, Christopher Seiler, Suresh Pujari, Kristopher Murphy, Shuo Yan, and Dan Murphy synthesized compounds on the way to making phosphoramidites. Susith Wickramaratne and Suresh Pujari took compound identification spectra including NMR. Christopher Seiler took HRMS of the synthesized molecules. Christopher Seiler synthesized the stable isotope labeled internal standard. Delshanee Kotandeniya, Christopher Seiler, and Lauren Curwick synthesized, purified, and characterized DNA oligos. Jenna Fernandez performed the computational modeling with direction from Yuk Y. Sham.

2.1 Introduction

5-Methylcytosine (MeC) is the most abundant endogenous modification of DNA known to date, with approximately 5% of all cytosine bases carrying the C-5 methyl group.[13] The majority of MeC residues are found within 5′-CG-3′ dinucleotides (CpG sites), where both cytidine nucleotides are methylated. Methylated CpG sites are recognized by methyl-CpG binding proteins (MBP),[226] which in turn recruit histone-modifying proteins, ultimately leading to chromatin remodeling and gene silencing.[11, 227]

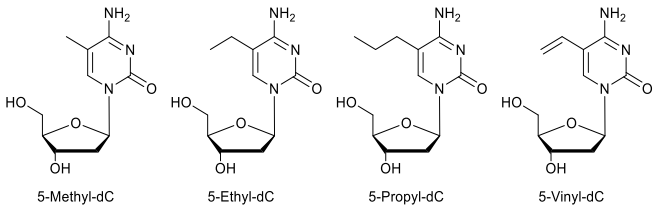
Methylation status of gene promoter regions largely determines which groups of genes are actively expressed in a given tissue.[228] Cytosine methylation patterns are stable in adult cells for many generations and are transmitted to progeny cells upon cell division.[229] The hemi-methylated 5′-CG-3′ sequences arising as a result of semiconservative replication are recognized by maintenance DNA methyltransferase (DNMT1), which places a methyl group on the C-5 position of cytosine of the newly synthesized strand opposite MeCG of the template strand, thereby preserving DNA methylation patterns.[11] Any errors in DNMT1 action would lead to aberrant gene expression, which in turn can contribute to malignant transformation (Chapter I of this thesis).[230]

It has been shown that DNMT1 fidelity can be compromised by structural modifications within 5′-CpG-3′ sites.[19, 231] For example, 5-halogenated cytosines mimic 5-methylcytosine, leading to erroneous methylation of CpG sites by

DNMT1.[19, 231-232] In contrast, oxidation of the C-5 methyl group of MeC to 5-hydroxymethyl-dC, 5-formyl-dC, and 5-carboxy-dC interferes with DNMT1 mediated methylation of the cytosine in the opposite strand, ultimately leading to the removal of DNA methylation marks (Chapter III of this thesis).[233]

The goal of the present work was to probe the ability of C5-alkylated cytosines with extended C5 side chain to direct DNMT1-catalyzed DNA methylation. The novel C5-alkyl-cytosine analogs (5-ethyl-dC, 5-propyl-dC, and 5-vinyl-dC, **Scheme 2.1**) were placed within a CpG site derived from the p53 tumor suppressor gene, and the resulting synthetic DNA duplexes were incubated with human recombinant DNMT1 protein. A mass spectrometry-based assay was used to study the kinetics of cytosine methylation in the presence of native MeC and its homologues with extended C5 alkyl chain. Furthermore, DNMT1 interactions with structurally modified DNA were investigated using molecular modeling. Our results indicate DNMT1 can the flexible, extended C5 side chain of 5-ethyl-dC but is unable to methylate DNA duplexes containing the more rigid 5-vinyl-dC or larger C5 side chain of 5-propyl-dC. Molecular modeling indicates the reduced activity toward the extended C5 alkyl side chains are due to a steric clash with the target recognition domain of the protein.

Scheme 2.1. Structures of MeC, ^{Ethyl}dC, ^{Propyl}dC and ^{Vinyl}dC.



2.2 Materials and Methods

Materials:

All nucleoside phosphoramidites, solvents, and solid supports required for the solid phase synthesis of DNA were acquired from Glen Research Corporation (Sterling, VA). Recombinant human DNA-methyltransferase 1 was purchased from New England Biolabs (Ipswich, MA). Phosphodiesterase I, phosphodiesterase II, and DNase I were obtained from Worthington Biochemical Corporation (Lakewood, NJ). Bovine intestinal alkaline phosphatase was procured from Sigma Aldrich Chemical Company (Milwaukee, WI) while 2′-deoxycytidine hydrochloride was from Berry & Associates Inc. (Dexter, MI). Stable isotope labeled thymidine was purchased from Cambridge Isotope Laboratories (Tewksbury, MA). The rest of the chemicals were purchased either from Fisher Scientific (Fairlawn, NJ), Sigma-Aldrich (Milwaukee, WI) or Alfa Aesar (WardHill, MA) and used without further purification unless noted otherwise. NMR was performed on a 500 MHz Bruker system. Accurate mass measurements were obtained on a LTQ Orbitrap Velos (Thermo Scientific, Pittsburgh, PA) and mass fragmentation data were collected on an Agilent 1100 LC/MSD Ion trap (Agilent Technologies, Santa Clara, CA).

Methods:

5-Iodo-2′-deoxycytidine (1) (**Scheme 2.2**)

2′-deoxycytidine hydrochloride, 4.0 g (15.2 mmol), was suspended in 3.0 mL of carbon tetrachloride, 4.5 mL of water, and 12.0 mL of glacial acetic acid. To the

suspension, was added I_2 2.3 g (9.06 mmol, 0.60 equiv) and iodic acid, 1.4 g (8.0 mmol, 0.53 equiv). The reaction mixture was stirred at 40 ^oC for 90 min after which 45 mL of water was added followed by cooling to room temperature. The mixture was filtered under vacuum, and the filtrate was extracted with benzene (10 x 100 mL) and reduced under vacuum resulting in a yellow residue. The residue was dissolved in 100 mL of methanol and evaporated under vacuum to remove excess acetic acid. This process was repeated 5 times to remove additional acetic acid, resulting in a yellow solid. The solid was dissolved in a minimum amount of deionized water and the pH adjusted to pH 10 with 10 N NaOH. This liquid was frozen at -20 ℃ overnight, warmed to a room temperature and then filtered. The filtrate was reduced to half its volume by heating, refrozen, warmed to room temperature, and filtered to yield a second crop of crystals. The crystals were combined and recrystallized in methanol to yield plate-like crystals of **1** (3.3 g, 61% yield). ESI^+-MS/MS: *m/z* 353.9 $[M + H^+]^+ \rightarrow$ *m/z* 238.0 $[M + H^+ - \text{deoxyribose}]^+$, Exact mass calc: 353.9945 $[M + H^+]^+$, Exact mass obs: 353.9939 $[M + H^+]^+$ (error 1.7 ppm); ^1H NMR (DMSO-d_6, 500 MHz): δ 8.28 (s, 1H), 7.79 (br s, 1H), 6.58 (br s, 1H), 6.07 (dd, 1H, *J* = 6.5, 6.5 Hz), 5.19 (d, 1H, *J* = 4.0 Hz), 5.09 (t, 1H, *J* = 5.0 Hz), 4.20 (ddd, 1H, *J* = 10.0, 4.0, 4.0 Hz), 3.78 (ddd, 1H, *J* = 6.5, 3.5, 3.5 Hz), 3.62 (ddd, 1H, *J* = 12.0, 5.0, 4.0 Hz), 3.54 (ddd, 1H, *J* = 12.0, 4.5, 4.0 Hz), 2.13 (ddd, 1H, *J* = 13.0, 6.0, 4.0 Hz), 1.99 (m, 1H). ^{13}C NMR (DMSO-*d6*, 125 MHz): δ 163.7, 153.9, 147.3, 87.4, 85.3, 69.9, 60.8, 56.5, 40.8.

5-iodo-3′,5′-*t*-butyldimethylsilyl-2′-deoxycytidine (2) (**Scheme 2.2**)

A flask containing compound **1,** 0.50 g (1.42 mmol), and imidazole, 0.77 g (11 mmol, 8 equiv), which was dried under high vacuum for 2 days. Anhydrous dimethylformamide (DMF) under vigorous stirring was bubbled with Ar for 30 min. 10 mL of the DMF was added to the nucleoside/imidazole mixture in an Ar-glove bag. To this mixture was added t-butyldimethylsilyl chloride (TBSCl), 0.85 g (5.7 mmol, 4 equiv), under an atmosphere of Ar. The reaction was stirred under Ar at room temperature for 48 h. The reaction was monitored by TLC (12:1 CH_2Cl_2:MeOH on silica plates). The reaction was reduced under vacuum and then taken up in CH_2Cl_2 (50 mL). The organic layer was washed with water (5 x 50 mL) and then dried using Na_2SO_4. The product was purified on a silica column eluted with 12:1 CH_2Cl_2:methanol to yield the final product **2** as a white solid (0.75 g, 91% yield). ESI^+-MS/MS: *m/z* 582.1 $[M + H^+]^+ \rightarrow$ *m/z* 237.9 $[M + H^+ -$ deoxyribose$]^+$, Exact mass calc: 582.1675 $[M + H^+]^+$, Exact mass obs: 582.1663 $[M + H^+]^+$ (error 2.1 ppm); ^{1}H NMR ($CDCl_3$, 500 MHz): δ 8.08 (s, 1H), 6.24 (dd, 1H, *J* = 6.5, 6.5 Hz), 4.20 (ddd, 1H, *J* = 6.0, 3.0, 3.0 Hz), 3.98 (d, 1H, *J* = 2.5 Hz), 3.88 (dd, 1H, *J* = 11.5, 2.5 Hz), 3.75 (dd, 1H, *J* = 11.5, 2.5 Hz), 2.45 (ddd, 1H, *J* = 13.5, 5.5, 3.0 Hz), 1.96 (m, 1H), 0.93 (s, 9H), 0.88 (s, 9H), 0.13 (d, 6H, *J* = 5.5 Hz), 0.06 (d, 6H, *J* = 3.5 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 163.7, 155.2, 146.8, 88.4, 87.0, 72.4, 63.0, 56.2, 42.8, 26.3, 25.9, 18.6, 18.1, -4.5, -4.7, -5.0, -5.1.

5-(Trimethylsilyl)-ethynyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (3) (**Scheme 2.3**)

A flask containing compound **2** (1.5 g, 2.6 mmol) was dried under vacuum for two days. The dried solid was transferred to a flame-dried flask and the solid was dissolved in 100 mL 3:1 dry, deoxygenated solution of DMF:Et_3N (deoxygenated by bubbling with Ar for 30 min). To the solution, 2 molar equivalents of trimethylsilylacetylene (0.73 mL, 5.2 mmol), 0.2 mol eq. of CuI (0.098 g, 0.52 mmol), and $Pd(PPh_3)_4$ (0.30 g, 0.26 mmol) were added sequentially to the flask in an Ar-glove bag. The flask was capped, and the reaction mixture was stirred for 2 days at room temperature under an atmosphere of Ar, under dark conditions. The mixture was diluted with 400 mL EtOAc and the organic layer was washed with water (100 mL x 3) and brine (twice) followed by drying the organic layer with anhydrous Na_2SO_4. The product was purified by two successive silica columns using a gradient of 100% CH_2Cl_2 to 16:1 CH_2Cl_2:2-propanol to give compound **3** in 35% yield. ESI^+-MS/MS: *m/z* 552.2 $[M + H]^+ \rightarrow$ *m/z* 208.1 $[M + H - \text{deoxyribose}]^+$, Exact mass calc: 552.3104 $[M + H]^+$, Exact mass obs: 552.3093 $[M + H]^+$ (error 2.0 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 8.23 (br s, 1H), 8.07 (s, 1H), 6.24 (dd, 1H, *J* = 6.0, 6.0 Hz), 5.76 (br s, 1H), 4.35 (ddd, 1H, *J* = 6.3, 3.3, 3.3 Hz), 3.95 (d, 1H, *J* = 2.5 Hz), 3.91 (dd, 1H, *J* = 11.5, 2.0 Hz), 3.75 (dd, 1H, *J* = 11.5, 2.0 Hz), 2.43 (ddd, 1H, *J* = 13.0, 6.0, 4.0 Hz), 2.00 (m, 1H), 0.92 (s, 9H), 0.87 (s, 9H), 0.21 (s, 9H), 0.13 (d, 6H, *J* = 6.5 Hz), 0.05 (d, 6H, *J* = 3.0 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 164.9, 154.4, 144.1, 101.2, 96.2, 91.1, 88.2, 86.8, 71.7, 62.6, 42.8, 26.2, 25.9, 18.6, 18.1, 0.0, -4.5, -4.8, -5.1, -5.3.

5-Ethynyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (4) (**Scheme 2.3**)

Compound **3** (0.7 g, 1.3 mmol) was dissolved in 10 mL anhydrous MeOH, and the solution was cooled on ice for 5 min. To this was added 3 mL of 1N KOH, and the mixture was stirred at room temperature for 3 hours. The reaction was monitored by TLC using 12:1 dichloromethane (DCM):MeOH. Once the starting material appeared to have been consumed, the pH of the solution was brought up to pH 7 by dropwise addition of 1N HCl. The solvent was then evaporated under reduced pressure, and the product was purified on a silica column (DCM:MeOH) to afford a light beige foam 0.55 g (90% yield). ESI^+-MS/MS: *m/z* 480.3 $[M + H]^+ \rightarrow$ *m/z* 136.0 $[M + H - \text{deoxyribose}]^+$, Exact mass calc: 480.2708 $[M + H]^+$, Exact mass obs: 480.2699 $[M + H]^+$ (error: 1.9 ppm); ^1H NMR ($CDCl_3$, 600 MHz): δ 8.21 (s, 1H), 6.26 (dd, 1H, *J* = 6.5, 6.5 Hz), 6.11 (br s, 1H), 5.63 (br s, 1H), 4.36 (m, 1H), 3.95 (m, 1H), 3.93 (dd, 1H, *J* = 9.7, 2.3 Hz), 3.77 (dd, 1H, *J* = 9.5, 1.5 Hz), 3.32 (s, 1H), 2.48 (m, 1H), 2.04 (ddd, 1H, *J* = 10.7, 5.3 Hz), 0.92 (s, 9H), 0.88 (s, 9H), 0.13 (d, 6H, *J* = 6.0 Hz), 0.06 (d, 6H, *J* = 3.0 Hz).

5-Ethyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (5) (**Scheme 2.3**)

Compound **4** (0.548 g, 1.14 mmol) was dried overnight under high vacuum and was dissolved in 20 mL of anhydrous MeOH by gently warming it in a water bath. The flask was capped, and the solution was purged with Ar by bubbling. To this was quickly added 0.35 mol equivalents of 10% Pd-C (0.042 g, 0.4 mmol) and the reaction was stirred under hydrogen for 24 h at room temperature. The reaction was monitored by TLC, using 12:1

DCM:MeOH as the mobile phase. The crude reaction mixture was then passed through a Celite® column to remove the Pd-catalyst, and the column washed with MeOH to collect the product. MeOH was evaporated under reduced pressure to yield 0.547 g of white foam (99% yield). ESI^+-MS/MS: *m/z* 484.3 $[M + H]^+ \rightarrow$ *m/z* 140.0 $[M + H - \text{deoxyribose}]^+$, Exact mass calc: 484.3021 $[M + H]^+$, Exact mass obs: 484.3008 $[M + H]^+$ (error 2.7 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 7.50 (s, 1H), 6.31 (dd, 1H, *J* = 6.5, 6.5 Hz), 4.34 (ddd, 1H, *J* = 6.0, 3.0, 3.0 Hz), 3.93 (d, 1H, *J* = 3.5 Hz), 3.83 (dd, 1H, *J* = 11.5, 3.5 Hz), 3.75 (dd, 1H, *J* = 11.5, 3.0 Hz), 2.40 (ddd, 1H, *J* = 13.0, 5.5, 3.0 Hz), 2.27 (q, 2H, *J* = 7.5 Hz), 1.93 (m, 1H), 1.15 (t, 3H, *J* = 7.5 Hz), 0.91 (s, 9H), 0.88 (s, 9H), 0.09 (s, 6H), 0.06 (d, 6H, *J* = 3.0 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 165.1, 155.9, 137.4, 107.2, 87.8, 86.1, 72.1, 63.1, 42.2, 26.1, 25.9, 21.0, 18.5, 18.1, 13.0, -4.5, -4.7, -5.3.

4-*N*-Benzoyl-5-ethyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (6) (Scheme 2.3)

Compound **5** (0.125 g, 0.258 mmol) which was dried overnight under high vacuum was co-evaporated three times with 5 mL each of anhydrous pyridine. The residue was dissolved in 5 mL of anhydrous pyridine, and the flask was immediately capped with a rubber septum under Ar atmosphere. To the mixture was added 60 µL (0.516 mmol, 2.0 equiv) of benzoyl chloride over a period of 5 minutes at room temperature and the reaction was stirred. The reaction was monitored by TLC (30:1 DCM:MeOH) every ½ hour. Since the reaction still had a lot of starting material remaining after 2 hours, another 40 µL (0.34 mmol, 1.3 equiv) of benzoyl chloride was added and reaction was monitored for another 2

hours, until all starting material was consumed. At the end of the reaction the solvent was removed under reduced pressure, the crude product was taken up in DCM, and the organic layer was washed three times with 50 mL water followed by drying with anhydrous Na_2SO_4 and the solvent removed under reduced pressure. The crude product was purified on a silica column, using 30:1 DCM:MeOH. After evaporation of solvent, a few drops of MeOH were added to the resulting oil to produce white needle like crystals 0.129 g (85% yield). ESI^+-MS/MS: Exact mass calc: 588.3284 $[M + H]^+$, Exact mass obs: 588.3267 $[M + H]^+$ (error 2.9 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 13.4 (s, 1H), 8.30 (dd, 2H, *J* = 8.5, 1.5 Hz), 7.62 (s, 1H), 7.52 (m, 1H), 7.44 (dd, 2H, *J* = 7.5, 7.5 Hz), 6.34 (dd, 1H, *J* = 7.5, 5.5 Hz), 4.41 (m, 1H), 3.98 (ddd, 1H, *J* = 5.0, 2.5 Hz), 3.98 (dd, 1H, *J* = 11.5, 3.0 Hz), 3.78 (dd, 1H, *J* = 11.5, 3.0 Hz), 2.57 (m, 2H), 2.34 (ddd, 1H, *J* = 13.5, 6.0, 2.5 Hz), 2.02 (ddd, 1H, *J* = 13.5, 7.5, 6.0 Hz), 1.25 (t, 3H, *J* = 7.5 Hz), 0.94 (s, 9H), 0.90 (s, 9H), 0.12 (d, 6H, *J* = 2.0 Hz), 0.09 (d, 6H, *J* = 4.0 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 179.7, 159.5, 147.9, 137.5, 136.4, 132.4, 130.0, 128.2, 117.5, 88.2, 85.7, 72.5, 63.2, 41.7, 26.1, 25.9, 21.6, 18.6, 18.1, 13.9, -4.5, -4.7, -5.3.

4-*N*-Benzoyl-5-ethyl-2′-deoxycytidine (7) (Scheme 2.3)

Compound **6** (0.046 g, 0.078 mmol) was dried overnight under high vacuum and then co-evaporated three times with 5 mL each of anhydrous pyridine. The flask was immediately fitted with a rubber septum and purged with Ar (1-2 min) and 1.3 mL of dry and degassed DCM was added under an atmosphere of Ar. To the solution, 102 µL (0.63

mmol, 8 equiv) of triethylamine-trihydrofluoride (Et_3N•3HF) was added dropwise to the flask over a period of 30 min using a syringe. The reaction was stirred under an atmosphere of Ar at room temperature for 16 h. During this time, the reaction was monitored every 2 h by TLC (30:1 DCM:MeOH, if a lot of starting material is remaining, more Et_3N.3HF is added). Upon total consumption of starting compound **6** by TLC, product (**7**) was isolated by purification on a silica column using a gradient of DCM and 30:1 DCM:MeOH resulting in a white solid 0.027 g in 96%, yield. ESI^+-MS/MS: *m/z* 360.2 $[M + H]^+ \rightarrow$ *m/z* 244.1 [M + H – deoxyribose]$^+$, Exact mass calc: 360.1554 $[M + H]^+$, Exact mass obs: 360.1553 $[M + H]^+$ (error 0.3 ppm); ^{1}H NMR (MeOD/DMSO-d_6, 500 MHz): δ 7.47 (d, 2H, *J* = 5.5 Hz), 7.33 (s, 1H), 6.78 (t, 1H, *J* = 7.3 Hz), 6.69 (t, 2H, *J* = 7.5 Hz), 5.52 (dd, 1H, *J* = 6.8, 6.8 Hz), 3.64 (m, 1H), 3.17 (d, 1H, *J* = 3.0 Hz), 3.03 (dd, 1H, *J* = 12.0, 3.0 Hz), 2.96 (dd, 1H, *J* = 11.8, 3.3 Hz), 1.81 (q, 2H, *J* = 7.5 Hz), 1.54 (m, 1H), 1.48 (m, 1H), 0.46 (t, 3H, *J* = 7.5 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 180.6, 160.7, 149.5, 139.4, 138.5, 133.6, 130.8, 129.3, 118.0, 89.2, 87.0, 71.9, 62.5, 41.7, 21.9, 13.7.

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-ethyl-2′-deoxycytidine (8) (**Scheme 2.3**)

Compound **7** (4-*N*-benzoyl-5-ethyl-2′-dC, 0.130 g, 0.361 mmol) was dried under vacuum overnight and co-evaporated three times with 6 mL each of anhydrous pyridine. The flask was immediately capped, purged with Ar, and transferred to a glove bag filled with Ar. Anhydrous pyridine (kept over activated 4Å molecular sieves) was degassed with Ar for 20 min. All additions were performed in an Ar glove bag. Anhydrous, degassed

pyridine (2 mL) was added to the flask containing dried nucleoside. Dimethoxytrityl chloride (DMTrCl), 0.147 g (0.434 mmol, 1.2 equiv), was added to the nucleoside-pyridine solution. Upon addition of all reagents, the flask was taken out of the glove bag, and the reaction was stirred for 16 h at room temperature away from light. The reaction was monitored using 5:47.5:47.5 TEA:EtOAc:Hexane on TLC plates soaked in 95:5 hexane:TEA. The reaction mixture was quickly dried under high vacuum. To this was added 30 mL of DCM and the organic layer was washed two times with 25 mL each of ice cold water. The reaction mixture was dried using anhydrous Na_2SO_4, and the solvent was removed under reduced pressure. The resulting solid was purified on a short silica-flash column packed in 95:5 Hexane:TEA using solvent gradient below: 95:5 Hexane:TEA, 5:25:70 TEA:EtOAc:Hexane, 5:47.5:47.5 TEA:EtOAc:Hexane, 95:5 EtOAc:TEA. Upon evaporation of the solvent, the resulting oily product was dissolved in 1 mL of anhydrous DCM and dried under high vacuum overnight to get a white foam 0.155 g of 65% yield. ESI^+-MS/MS: *m/z* 662.6 $[M + H]^+ \rightarrow$ *m/z* 303.1 [dimethoxytrityl cation]$^+$, 244.0 [M + H – deoxyribose]$^+$, Exact mass calc: 684.2680 $[M + Na]^+$, Exact mass obs: 684.2680 $[M + Na]^+$; ^1H NMR ($CDCl_3$, 500 MHz): δ 8.20 (dd, 1H, *J* = 7.0, 1.0 Hz), 7.64 (s, 1H), 7.51 (m, 1H), 7.42 (m, 4H), 7.30 (m, 7H), 6.84 (m, 4H), 6.40 (dd, 1H, *J* = 6.0, 5.0 Hz), 4.57 (ddd, 1H, *J* = 5.0, 2.5, 2.5 Hz), 4.07 (ddd, 1H, *J* = 3.0, 3.0, 3.0 Hz), 3.50 (dd, 1H, *J* = 9.0, 3.0 Hz), 3.39 (dd, 1H, *J* = 8.7, 2.7 Hz), 2.60 (br s, 1H), 2.47 (dq, 1H, *J* = 11.0, 2.5 Hz), 2.32 (m, 1H), 2.23 (m, 1H), 2.09 (m, 1H), 1.00 (t, 3H, *J* = 6.3 Hz).

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-ethyl-2-deoxycytidine-3′-[(2-cyanoethyl)-(*N,N*-diisopropyl)]-phosphoramidite (9) (**Scheme 2.3**)

Into a flame dried round bottom flask was added compound **8** (0.075 g, 0.113 mmol), which was dried in an Ar filled glove bag under high vacuum overnight. Anhydrous DCM (kept over activated 4Å molecular sieves) was degassed with argon for 20 min. All additions below were performed sequentially in an Ar glove bag: 197 µL of anhydrous *N,N*-diisopropylethylamine (1.13 mmol, 5 equiv as phosphitylating reagent), 2 mL of anhydrous and degassed DCM and 51 µL of 2-Cyanoethyl *N,N*-diisopropylchlorophosphoramidite (0.23 mmol, 2 equiv). The flask was capped, and the reaction stirred for 16 h under an Ar atmosphere at room temperature away from light. After 16 h, 250 µL of anhydrous MeOH was added to prevent the oxidation of the phosphoramidite and 200 µL of Et_3N was added to ensure that the product does not get hydrolyzed during the work-up steps. To this solution was added 30 mL of EtOAc. The organic layer was extracted three times with ice-cold brine, dried using anhydrous Na_2SO_4, and the solvent was removed under reduced pressure. The resulting solid was dissolved in a few mL of 5:95 TEA:EtOAc and a short silica-flash column was run to isolate the product using the gradient elution described below: 95:5 Hexane:TEA, 5:25:70 TEA:EtOAc:Hexane, 5:47.5:47.5 TEA:EtOAc:hexane, 95:5 EtOAc:TEA. Evaporation of the solvent resulted in compound 8 as a white foam (0.059 g, 62% yield). ^{1}H NMR ($CDCl_3$, 400 MHz): δ 8.28 (d, 2H, J = 7.6 Hz), 7.67 (s, 1H), 7.51 (t, 1H, 8.0 Hz), 7.42-7.27 (m, 11H), 6.85 (m, 4H), 6.41 (dd, 1H, J = 13.8, 7.6 Hz), 4.66 (m, 1H), 4.18 (m, 1H), 3.79 (s,

6H), 3.63-3.48 (m, 4H), 3.35-3.30 (m, 2H), 2.62 (t, 2H, J = 6.4 Hz), 2.41 (dd, 1H, J = 6.4, 6.4 Hz), 2.31 (m, 1H), 2.17 (m, 1H), 1.25-1.16 (m, 12H), 0.96 (t, 3H, J = 8.4 Hz). ^{31}P NMR ($CDCl_3$, 162 MHz): δ 147.70, 147.29.

5-Propyn-1-yl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (10) (**Scheme 2.4**)

5-Iodo-3′,5′-TBS-2′-dC (**2**) (2 g, 3.44 mmol) was dried under vacuum for two days. The dried solid was transferred to a flame-dried two-necked flask inside an Ar-glove bag. To the flask was added 0.397 g of $Pd(PPh_3)_4$ (0.344 mmol, 0.1 equiv) and 0.033 g CuI (0.172 mmol, 0.05 equiv). The flask was capped, and 50 mL of anhydrous, deoxygenated solution of DMF:Et_3N (4:1) was added, along with constant purging of propyne gas. The pressure inside the flask was kept at a constant of 8 psi throughout the reaction. The reaction was stirred for 48 h at room temperature, away from light. At the end of the reaction, the solvents were evaporated, and crude product was taken up in 150 mL of DCM. The organic layer was washed five times with 75 mL each of water and twice with 50 mL each of brine, followed by drying organic layer with anhydrous Na_2SO_4. The product was purified by two successive silica columns with solvent gradient from 100% DCM to 9:1 DCM:MeOH to afford light a beige solid in (1.44 g) 85% yield. ESI$^+$-MS/MS: *m/z* 494.3 $[M + H]^+ \rightarrow$ *m/z* 150.1 $[M + H - \text{deoxyribose}]^+$, Exact mass calc: 494.2865 $[M + H]^+$, Exact mass obs: 494.2849 $[M + H]^+$ (error 3.2 ppm); ^{1}H NMR ($CDCl_3$, 500 MHz): δ 8.44 (br s, 1H), 8.00 (s, 1H), 6.25 (dd, 1H, J = 6.0, 6.0 Hz), 5.81 (br s, 1H), 4.34 (m, 1H), 3.90 (m, 2H), 3.74 (dd, 1H, J = 12.0, 3.0 Hz), 2.40 (ddd, 1H, J = 13.5, 6.5, 4.5 Hz), 2.00 (m, 4H),

0.92 (s, 9H), 0.86 (s, 9H), 0.12 (d, 6H, J = 8.0 Hz), 0.04 (s, 6H). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 165.2, 154.5, 142.9, 91.8, 91.7, 87.9, 86.4, 71.3, 71.1, 62.5, 42.6, 26.1, 25.8, 18.5, 18.1, 4.5, -4.5, -4.8, -5.4, -5.5.

5-Propyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (11) (**Scheme 2.4**)

To a flask containing 0.075 g of dry 5-propyn-1-yl-3′,5′-*t*-butyldimethylsilyl-2′-deoxycytidine (**10**) was added 5 mL of anhydrous MeOH. The flask was capped, and the solution was purged with Ar by bubbling for 5 min. To this was added 20 weight % of 10% Pd-C catalyst (0.015 g, 0.14 mmol). The reaction mixture was kept under hydrogen gas and stirring at room temperature, away from light for 2 days. The reaction was monitored by TLC, using 12:1 DCM:MeOH as the mobile phase. The crude reaction was then passed through a Celite® column to remove the Pd catalyst, and the column washed with MeOH to collect product **11**. The solvent was evaporated under reduced pressure to yield white crystals (0.068 g) in 90% yield. ESI$^+$-MS/MS: *m/z* 498.2 $[M + H]^+ \rightarrow$ *m/z* 154.1 $[M + H - \text{deoxyribose}]^+$, Exact mass calc: 498.3178 $[M + H]^+$, Exact mass obs: 498.3163 $[M + H]^+$ (error 3.0 ppm); ^{1}H NMR ($CDCl_3$, 500 MHz): δ 7.46 (s, 1H), 6.30 (dd, 1H, J = 6.8, 6.8 Hz), 4.33 (m, 1H), 3.91 (ddd, 1H, J = 6.0, 3.0, 3.0 Hz), 3.82 (dd, 1H, J = 11.0, 3.0 Hz), 3.74 (dd, 1H, J = 11.0, 3.0 Hz), 2.37 (ddd, 1H, J = 13.0, 6.0, 3.0 Hz), 2.20 (t, 2H, J = 7.5 Hz), 1.91 (m, 1H), 1.51 (qt, 2H, J = 7.5, 7.5 Hz), 0.92 (t, 3H, J = 7.5 Hz), 0.90 (s, 9H), 0.86 (s, 9H), 0.08 (d, 6H, J = 2.0 Hz), 0.04 (d, 6H, J = 2.5 Hz). ^{13}C NMR ($CDCl_3$, 125

MHz): δ 165.3, 155.8, 137.9, 105.9, 87.7, 85.9, 72.0, 63.0, 42.2, 30.0, 26.1, 25.9, 21.8, 18.5, 18.1, 13.8, -4.5, -4.8, -5.3.

4-*N*-Benzoyl-5-propyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (12) (**Scheme 2.4**)

Compound **11** (0.2 g, 0.4 mmol) was co-evaporated three times with anhydrous pyridine (5 mL), and the flask was flushed with Ar gas. To this was added 6 mL of anhydrous pyridine, followed by purging of the solution with argon by for 5 min. To the flask, 56 µL of benzoyl chloride (0.48 mmol, 1.2 equiv) was added by syringe over 5 min. This mixture was stirred at room temperature under an argon atmosphere for 4 h. The course of the reaction was monitored by TLC (30:1 DCM:MeOH). Analysis of the crude mixture by mass spectrometry indicated the dibenzoylated product to be present. The crude product mixture was dissolved in 2.5 mL of 65:30:5 pyridine:MeOH:H_2O, and this mixture was cooled on ice. To this was added 2.5 mL of cold 2N NaOH solution, in 65:30:5 pyridine:MeOH:H_2O. The reaction mixture was stirred on ice for 10 min. Ammonium chloride (0.1 g, 1.2 equiv) was added to quench the reaction. The crude sample was dried and re-dissolved in DCM, washed three times with 30 mL water, and dried with Na_2SO_4, and the organic layer was removed to result in crude solid **12**. Compound **12** was further purified by silica column using 15:1 DCM:MeOH to give colorless solid (0.217 g) in 90% yield. ESI^+-MS/MS: Exact mass calc: 602.3440 $[M + H]^+$, Exact mass obs: 602.3421 $[M + H]^+$ (error 3.2 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 8.29 (dd, 2H, J = 7.0, 1.5 Hz), 7.62 (s, 1H), 7.52 (t, 1H, J = 7.5 Hz), 7.44 (t, 2H, J = 7.5 Hz), 6.35 (dd, 1H, J = 7.5, 5.5 Hz),

4.41 (m, 1H), 3.97 (ddd, 1H, *J* = 3.0, 3.0, 3.0 Hz), 3.88 (dd, 1H, *J* = 11.3, 2.8 Hz), 3.78 (dd, 1H, *J* = 11.3, 2.8 Hz), 2.51 (m, 2H), 2.33 (ddd, 1H, *J* = 13.0, 6.0, 2.5 Hz), 2.02 (ddd, 1H, *J* = 13.5, 7.5, 6.0 Hz), 1.68 (qt, 2H, *J* = 6.5, 6.5 Hz), 1.00 (t, 3H, *J* = 7.5 Hz), 0.94 (s, 9H), 0.90 (s, 9H), 0.13 (d, 6H, *J* = 2.5 Hz), 0.09 (d, 6H, *J* = 3.5 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 179.7, 159.7, 148.0, 137.5, 136.9, 132.5, 130.0, 128.2, 116.0, 88.2, 85.6, 72.4, 63.2, 41.8, 30.2, 26.1, 25.9, 22.5, 18.6, 18.1, 14.1, -4.5, -4.7, -5.3.

4-*N*-Benzoyl-5-propyl-2′-deoxycytidine (13) (**Scheme 2.4**)

Compound **12** (0.25 g, 0.502 mmol) was dried under vacuum and co-evaporated with of anhydrous pyridine (3 x 5 mL), and the flask was purged with Ar. To this was added 7 mL of anhydrous DCM to dissolve the solid, followed by purging of the solution mixture with Ar for 5 min. To the flask, was injected 327 µL of $Et_3N{:}H_3F$ (2.01 mmol) by syringe over 30 minutes. The reaction was stirred, away from light, for 16 h at room temperature. The course of the reaction was monitored by TLC (30:1, DCM:MeOH). The resulting crude product was purified using column chromatography with a mobile phase of 15:1 DCM:MeOH to afford compound **13** (0.131 g) as a white solid in 92% yield. ESI$^+$-MS/MS: *m/z* 374.0 $[M + H]^+ \rightarrow$ *m/z* 258.0 [M + H – deoxyribose]$^+$, Exact mass calc: 374.1710 $[M + H]^+$, Exact mass obs: 374.1704 $[M + H]^+$ (error 1.6 ppm); ^{1}H NMR (MeOD/DMSO-*d6*, 500 MHz): δ 8.49 (d, 2H, *J* = 6.0 Hz), 7.38 (s, 1H), 6.82 (t, 1H, *J* = 7.3 Hz), 6.73 (t, 2H, *J* = 7.5 Hz), 5.53 (dd, 1H, *J* = 6.5, 6.5 Hz), 3.64 (m, 1H), 3.19 (ddd, 1H, *J* = 3.0, 3.0, 3.0 Hz), 3.03 (dd, 1H, *J* = 11.7, 3.3 Hz), 2.96 (dd, 1H, *J* = 12.0, 3.5 Hz), 1.77

(m, 2H), 1.55 (m, 1H), 1.49 (m, 1H), 0.92 (m, 2H), 0.24 (t, 3H, J = 7.5 Hz). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 180.3, 160.8, 149.3, 140.2, 138.6, 133.7, 130.8, 129.5, 116.0, 89.3, 86.9, 71.8, 62.5, 41.7, 30.5, 22.8, 14.4.

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-propyl-2′-deoxycytidine (14) (**Scheme 2.4**)

Compound **13** (50 mg, 0.13 mmol) was dried under vacuum overnight and co-evaporated with anhydrous pyridine (3 x 5 mL) and this flask was kept under vacuum overnight. The flask was immediately capped and purged with Ar and transferred to a glove bag filled with Ar. Anhydrous pyridine (kept over activated 4Å molecular sieves) was degassed with Ar for 20 min. To the solution in the glove bag, 3 mL of anhydrous, degassed pyridine solution and dimethoxytrityl chloride (DMTrCl) (0.091 g, 0.27 mmol) were added sequentially. Upon addition of all reagents, the flask was taken out of the glove bag and the reaction was stirred for 16 h at room temperature away from light. The progress of the reaction was monitored by TLC. After completion of the reaction, the crude sample was quickly dried under high vacuum and 30 mL of DCM was added. The organic layer was washed three times with 25 mL each of ice cold H_2O, dried using anhydrous Na_2SO_4, and the solvent removed under reduced pressure. The resulting solid was purified on a short silica-flash column packed in 95:5 Hexane:TEA using the solvent gradient: 95:5 Hexane:TEA, 5:25:70 TEA:EtOAc:Hexane, 5:47.5:47.5 TEA:EtOAc:Hexane, 95:5 EtOAc:TEA. Upon evaporation of solvent, the resulting oily product was dissolved in 1 mL of anhydrous DCM and allowed to dry under high vacuum overnight, to get compound

14 as a white foam (0.075 g, 77% yield). ESI^+-MS/MS: *m/z* 676.1 $[M + H]^+ \rightarrow$ *m/z* 303.1 [dimethoxytrityl cation]$^+$, 258.1 [M + H – deoxyribose]$^+$, Exact mass calc: 698.2837 [M + Na]$^+$, Exact mass obs: 698.2833 [M + Na]$^+$ (error 0.5 ppm); ^{1}H NMR (ACN-*d*4, 400 MHz): δ 8.27 (dd, 2H, *J* = 6.8, 1.6 Hz), 7.69 (s, 1H), 7.59 (m, 1H), 7.49 (m, 4H), 7.34 (m, 7H), 6.90 (m, 4H), 6.27 (dd, 1H, *J* = 6.6, 6.6 Hz), 4.50 (ddd, 1H, *J* = 5.2, 5.2, 5.2 Hz), 4.00 (ddd, 1H, *J* = 3.6, 3.6, 3.6 Hz), 3.79 (s, 6H), 3.33 (d, 2H, *J* = 3.6 Hz), 2.35 (t, 2H, *J* = 6.4 Hz), 2.16 (m, 2H), 1.47 (m, 2H), 0.80 (t, 3H, *J* = 7.4 Hz).

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-propyl-2′-deoxycytidine-3′-[(2-cyanoethyl)-(*N*,*N*-diisopropyl)]-phosphoramidite (15) (**Scheme 2.4**)

Into a flame dried round bottom flask was added 4-*N*-Benzoyl-5-*O*′-(dimethoxytrityl)-5-propyl-2′-deoxycytidine (**14**, 0.050 g, 0.074 mmol) that was dried under high vacuum overnight, in an Ar filled glove bag. Anhydrous DCM (kept over activated 4 Å molecular sieves) was degassed with Ar for 20 min. All additions were performed in an Ar glove bag in the following order: anhydrous diisopropylethylamine (129 µL, 0.074 mmol), anhydrous DCM (2.5 mL) and 2-Cyanoethyl *N,N*-diisopropylchlorophosphoramidite (33 µL, 0.148 mmol). The flask was capped, and the reaction mixture stirred for 15 h under an argon atmosphere at room temperature away from light. After 15 h, 250 µL of anhydrous MeOH was added to prevent the oxidation of phosphoramidite and 200 µL of Et_3N was added to prevent hydrolysis during work-up. To this solution was added 30 mL of EtOAc and the organic layer extracted three times with

ice-cold brine, dried over anhydrous Na_2SO_4 and dried under reduced pressure. The resulting solid was dissolved in a 5 mL of 5:95 TEA:EtOAc and purified on a short silica-flash column using the gradient elution described below: 95:5 hexane:TEA, 5:25:70 TEA:EtOAc:hexane, 5:47.5:47.5 TEA:EtOAc:hexane, 95:5 EtOAc:TEA. Evaporation of solvent resulted in a white foam (0.049 g, 76% yield). ^{1}H NMR (Acetone-d_6, 400 MHz): δ 8.26 (d, 2H, J = 7.2 Hz), 7.82 (s, 1H), 7.57-7.44 (m, 5H), 7.39-7.23 (m, 7H), 6.91 (m, 4H), 6.35 (m, 1H), 4.77 (m, 1H), 4.21 (m, 1H), 3.90-3.80 (m, 1H), 3.77 (s, 6H), 3.75-3.58 (m, 3H), 3.47 (m, 2H), 2.76 (m, 2H), 2.62 (t, 2H, J = 6.0 Hz), 2.56 (m, 1H), 2.26 (m, 1H), 1.48 (m, 2H), 1.19-1.07 (m, 12H), 0.76 (t, 3H, J = 7.2 Hz). ^{31}P NMR (ACN-d4, 162 MHz): δ 148.37, 148.21.

3′,5′-*O*-*t*-butyldimethylsilyl-5-vinyl-2′-deoxycytidine (16) (**Scheme 2.5**)

Compound **2** (2.5 g, 4.3 mmol) was dried under high vacuum for 2 days and then coevaporated three times with 30 mL each anhydrous pyridine, and then dried under high vacuum for another two days. The flask was taken into an Ar-glove bag and to the nucleoside was added 75 mL of anhydrous deoxygenated 1-methyl-2-pyrolidinone. The resulting solution was again degassed by bubbling Ar for 2 h. The remaining additions were done inside an Ar-glove bag. To the nucleoside was added tri(2-furanyl)phosphine (0.55 g, 2.35 mmol), $Pd_2(dba)_3$ (0.504 g, 0.55 mmol) and tributylvinyltin (1.9 mL, 6.5 mmol) sequentially. The flask was sealed, and the reaction mixture was stirred, away from light for 1 day at 60 °C. The reaction was monitored by TLC with 16:1 DCM:IPA and 9:1

EtOAC:hexane. Since a little starting material remained, another portion of tributylvinyltin (0.8 mL, 3.25 mmol) was added, and the reaction mixture was stirred for another 12 h. The crude reaction containing the product was divided to five portions and each portion diluted with 100 mL of EtOAC and washed four times with 25 mL each H_2O. The combined organic layers were dried with Na_2SO_4 and concentrated under reduced pressure. The resulting crude product was dissolved in 200 mL acetonitrile and washed five times with 25 mL each of hexane. The ACN layer was partitioned over hexane and the ACN was dried to afford crude compound **16**. The product was purified on a silica column using a gradient of DCM to 15:85 IPA:DCM to obtain compound **16** (1.2 g) in 58% yield. ESI^+-MS: Exact mass calc: 482.2865 $[M + H]^+$, Exact mass obs: 482.2850 $[M + H]^+$ (error 3.1 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 7.72 (s, 1H), 6.32 (dd, 1H, *J* = 17.0, 11.0 Hz), 6.26 (dd, 1H, *J* = 6.3, 6.3 Hz), 5.40 (d, 1H, *J* = 17.5 Hz), 5.24 (d, 1H, *J* = 11.0 Hz), 4.33 (d, 1H, *J* = 2.5 Hz), 3.94 (s, 1H), 3.83 (d, 1H, *J* = 11.0 Hz), 3.73 (d, 1H, *J* = 11.5 Hz), 2.42 (m, 1H), 1.94 (m, 1H), 0.86 (s, 18H), 0.04 (d, 6H, *J* = 4.5 Hz), 0.03 (s, 6H). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 164.1, 155.3, 138.2, 128.4, 117.2, 105.8, 88.0, 86.4, 72.1, 63.0, 42.4 26.0, 25.8, 18.5, 18.1, -4.5, -4.8, -5.3, -5.4.

4-*N*-Benzoyl-3′,5′-*O*-(*t*-butyldimethylsilyl)- 5-vinyl-2′-deoxycytidine (17) (Scheme 2.5)

Compound **16** (0.81 g, 1.68 mmol) was dried on high vacuum overnight, co-evaporated with anhydrous pyridine (3 x 5 mL) and dissolved in 20 mL of anhydrous

pyridine. The flask was capped with a rubber septum under Ar atmosphere. To the stirred solution was added benzoyl chloride (234 µL, 2.01 mmol) and stirred for overnight at room temperature. The reaction was checked by TLC (30:1 DCM:MeOH). The next morning, another 234 µL (2.01 mmol) of benzoyl chloride was added, and the reaction was stirred for 24 h. Solvents were evaporated under reduced pressure, and the residue was taken up in 30 mL of DCM, which was washed with water (3 x 30 mL) and dried over sodium sulfate for 30 minutes. The solution was filtered, evaporated under reduced pressure, and subjected to flash column. Purification was achieved using a gradient of DCM:MeOH to give compound **17** as a colorless solid (0.807 g, 82% yield). ESI^+-MS: Exact mass calc: 586.3127 $[M + H]^+$, Exact mass obs: 586.3137 $[M + H]^+$ (error 1.7 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 13.13 (s, 1H), 8.16-8.17 (d, 2H, *J* = 7.5 Hz), 7.99 (S, 1H), 7.58-7.61 (m, 1H), 7.49-7.52 (m, 2H), 6.77-6.80 (m, 1H), 6.13 (t, 1H), 5.93-5.96 (d, 1H, *J* = 6.5 Hz), 5.32-5.34 (d, 1H, *J* = 17.5 Hz), 4.39-4.40 (m, 1H), 3.80-3.90 (m, 1H), 3.81-3.84 (m, 1H), 3.74-3.77 (m, 1H), 2.36-2.42 (m, 1H), 2.21-2.27 (m, 1H), 0.87-0.88 (2s, 20H), -0.12--0.10 (2s, 12H). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 141.32, 139.24, 135.35, 132.04, 131.10, 130.88, 118.07, 114.08, 90.21, 88.65, 74.57, 65.26, 28.49, 28.35, 20.75, 20.48, 20.39, -0.50, -2.07, -2.24, -2.74.

4-*N*-Benzoyl-5-vinyl-2′-deoxycytidine (18) (Scheme 2.5)

Compound **17** (0.81 g, 1.38 mmol) was dried on high vacuum overnight then co-evaporated three times with 5 mL of anhydrous pyridine. To this was added 15 mL

anhydrous DCM. The solution and flask were purged with argon for 5 min. To the flask, 900 µL (5.5 mmol) triethylamine trihydrofluoride was added over 30 minutes. The reaction was stirred in the dark for 16 h at room temperature. The completion of the reaction was monitored by TLC. The reaction mixture was evaporated under reduced pressure and the residue was extracted with DCM and dried over anhydrous Na_2SO_4 and purified on a silica gel column run washed with 15:1 DCM:MeOH and eluted with a mixture of 9:1 DCM:MeOH. Yield (0.64 g, 80%). ESI^+-MS: Exact mass calc: 358.1397 $[M + H]^+$, Exact mass obs: 358.1398 $[M + H]^+$ (error 0.3 ppm); ^1H NMR ($CDCl_3$, 500 MHz): δ 8.64 (s, 1H), 8.16-8.18 (d, 2H, *J* = 7.0 Hz), 7.59 (t, 1H), 7.52 (t, 2H), 6.77-6.82 (m, 1H), 6.17 (t, 1H), 5.84-5.88 (d, 1H, *J* = 18 Hz), 5.26-5.31 (m, 3H), 4.30-4.32 (m, 1H), 3.87-3.88 (m, 1H), 3.71-3.75 (m, 1H), 3.62-3.66 (m, 1H), 2.27 (t, 2H). ^{13}C NMR ($CDCl_3$, 125 MHz): δ 141.32, 139.24, 135.35, 132.04, 131.10, 130.88, 118.07, 114.08, 90.21, 88.65, 74.57, 65.26, 28.49, 28.35.

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-vinyl-2′deoxycytidine (19) (**Scheme 2.5**)

Compound **18** (155 mg, 0.43 mmol) was dried by repeated co-evaporation with dry pyridine (3 × 5 mL). The residue was dissolved in dry pyridine (6 mL) and 4,4′-dimethoxytrityl chloride (220 mg, 0.64 mmol) was added in portions and stirred at rt for 7h. After completion of the reaction (monitored by TLC), methanol (1 mL) was added to the reaction mixture, and stirring was continued for another 30 min. The reaction mixture was evaporated to dryness under reduced pressure, and the remaining residue was dissolved

in dichloromethane (25 mL) and washed with 5% aq. $NaHCO_3$ solution (50 mL) and water (50 mL). The organic layer was dried over Na_2SO_4, the solvent was evaporated under reduced pressure, and the residue was subjected to flash chromatography (silica gel, column 10 × 4 cm, eluted with CH_2Cl_2/acetone 95:5 to 85:15). Evaporation of the fractions corresponding to the product on TLC afforded **19** (0.167 g, 60%) as a pale-yellow foam. TLC (silica gel, CH_2Cl_2/MeOH 9:1). R_f 0.8. Exact mass calc: 660.2704 $[M + H]^+$, Exact mass obs: 660.2688 $[M + H]^+$ (error 2.4 ppm). ^{1}H- NMR [DMSO (d_6), 500 MHz]: δ 2.26-2.29 (m, 1H), 2.37-2.40 (m, 2H), 3.25-2.52 (d, 1H), 3.72 (s, 3H), 3.73 (s, 3H), 3.98-3.99 (d, 1H), 4.28 (s, 1H), 5.08-5.10 (d, 1H), 5.37-5.38 (d, 1H), 5.70-5.74 (d, 1H), 6.19 (t, 1H), 6.44-6.50 (m, 1H), 6.87-6.90 (m, 4H), 7.21-7.27 (m, 5H), 7.31 (t, 2H), 7.38-7.40 (d, 2H), 7.51 (t, 2H), 7.60 (t, 1H), 7.96 (brs, 1H), 8.16-8.17 (d, 2H). ^{13}C- NMR [DMSO (d_6), 500 MHz]: δ 173.75, 158.61, 158. 60, 157.70, 147.25, 145.13, 139.24, 137.03, 135.86, 135.83, 133.11, 130.20, 130.17, 129.81, 128.87, 128.37, 128.16, 127.24, 115.84, 113.70, 111.90, 86.26, 70.61, 64.07, 55.49, 55.48, 31.15.

4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-vinyl-2′deoxycytidine 3′-(2-cyanoethyl)-*N*, *N*′-diisopropyl phosphoramidite (20) (**Scheme 2.5**)

A solution of **19** (100 mg, 0.15 mmol) in dry CH_2Cl_2 (10 mL) was stirred with (*i*-Pr)$_2$NEt (52 μL, 39 mg, 0.30 mmol) at room temperature. Then, 2-cyanoethyl *N,N*-diisopropylchlorophosphoramidite (67 μL, 72 mg, 0.30 mmol) was added and the reaction mixture was stirred for 30 min. Upon completion of the reaction (monitored by TLC), the

reaction mixture was diluted with CH_2Cl_2 (30 mL) and was poured into 5% $NaHCO_3$ solution (30 mL) and extracted with CH_2Cl_2 (3 × 20 mL). The combined organic phases were dried over Na_2SO_4, and the solvent was evaporated. The residual foam was applied to a flash chromatography column (silica gel, 8 × 3 cm, eluted with CH_2Cl_2/acetone 100:0 → 95:5). Evaporation of the fractions corresponding to the product on TLC afforded **20** (90 mg, 70%) as a colorless viscous oil. TLC (silica gel, CH_2Cl_2/acetone 9:1). R_f 0.7. ESI^+-MS/MS: Exact mass calc: 860.3783 $[M + H]^+$, Exact mass obs: 860.3771 $[M + H]^+$ (error 1.4 ppm). ^{31}P NMR ($CDCl_3$, 500 MHz): 149.24, 148.64.

$^{13}C_{10}{}^{15}N_2$-5-Methyl-2′-deoxycytidine (21)

$^{13}C_{10}{}^{15}N_2$-thymidine (3.5 mg, 0.014 mmol, Cambridge Isotope Labs, Tewksbury MA) was dissolved in 0.7 mL DCM. To the stirred solution, 360 µL triethylamine (TEA) (2.6 mmol) and 270 µL chlorotrimethyl-silane (2.2 mmol) were added. The solution was stirred for 30 min at 23 °C and then quenched with 2 mL cold 1 M sodium bicarbonate ($NaHCO_3$). Phases were separated, the organic layer dried with anhydrous magnesium sulfate ($MgSO_4$), filtered, and evaporated under reduced pressure. The resulting brown residue was co-evaporated with anhydrous toluene (3 x 1 mL). At the same time, a suspension of 1,2,4-triazole (120 mg, 1.7 mmol), TEA (370 µL, 2.7 mmol), and phosphorous oxychloride ($POCl_3$) (37 µL, 0.40 mmol) in 2 mL acetonitrile (ACN) was stirred on ice for 30 min. The residue was then dissolved in 3 mL ACN, added to the suspension of triazole, TEA, and $POCl_3$, and stirred for 2 h at room temperature. The

reaction mixture was concentrated under reduced pressure and re-dissolved in 2 mL of DCM. The DCM solution was washed with 1 M cold $NaHCO_3$, separated, dried with $MgSO_4$, filtered, and the solvent was removed under vacuum. The residue was dissolved in 5.5 mL of 7 N methanolic ammonia and placed with a magnetic stir bar in a Teflon tube, which had been chilled on ice for 30 min. The Teflon tube was capped and placed in a stainless-steel bomb that was then sealed. The bomb was placed in an oil bath preheated to 75 °C and allowed to react for 3 days. After cooling the stainless-steel bomb to below room temperature, the bomb was opened, and the remaining ammonia was allowed to evaporate overnight. The residue was dissolved in 50:50 MeOH:DCM and dried prior to HPLC purification on an Agilent 1100 system. Purification was done using a Synergi 4u Hydro-RP 80A column [250 mm x 4.6 mm] from Phenomenex (Torrance, CA). The starting material and product were eluted isocratically using 6% MeOH in 10 mM ammonium formate at pH 4.2 for 32 min followed by an increase to 50% ACN and subsequent equilibration to initial conditions. The product **21** eluted at 12.5 min and starting material at 20.9 min. Carbograph SPE cartridges (Grace Analytical, Deerfield, IL) were conditioned in MeOH and then water, the product was loaded onto the cartridge, followed by a water wash and 3 successive methanol elutions. Product identity was confirmed by matching the UV spectra to a commercial unlabeled standard and HPLC-ESI^+-MS/MS. ESI^+-MS/MS: *m/z* 253.9 $[M + H]^+ \rightarrow$ *m/z* 133.2 [M + H - deoxyribose]$^+$ Exact mass calc: 254.1412 $[M + H]^+$, Exact mass obs: 254.1408 $[M + H]^+$ (error 1.6 ppm).

Solid Phase Synthesis and Purification of DNA Oligodeoxynucleotides

DNA sequences used in this study were derived from codons 153-158 of the human *p53* tumor suppressor gene (5′-CCCGGCACC CGC[$^{15}N_3$,$^{13}C_1$-G]TCCGCG-3′) (**Table 2.1**). Oligodeoxynucleotides containing a structurally modified C:G base pair at the first position of *p53* codon 157 were synthesized using an ABI 394 DNA Synthesizer (Applied Biosystems, California) in accordance to standard solid phase oligodeoxynucleotide synthesis protocols.[234]

The oligodeoxynucleotides were purified by reversed phase HPLC using an Agilent 1100 system interfaced with a UV-VWD detector.[235-236] The oligodeoxynucleotides containing modified cytosine base were purified on a Supelcosil LC-18-DB column (10 mm × 250 mm, 5 μm, Supelco, Bellefonte, PA) eluted at 40°C and a flow rate of 3 mL/min. HPLC buffers were 100 mM triethylammonium acetate, pH 7.0 (A), and acetonitrile (B). A linear gradient of 16.8-26% B in 21 minutes followed by a gradient increase to 34% B in the next 19 minutes was used. Following HPLC purification, the oligodeoxynucleotides were desalted by NAP-5-illustra™ size exclusion columns (GE Healthcare, Piscataway, NJ) according to the manufacturer's protocols. The presence of C5-alkyl-dC in the oligodeoxynucleotides was confirmed by capillary HPLC-ESI$^-$-MS (**Table 2.1**),[236-238] and their concentrations were determined from the extent of 2'-deoxyguanosine in enzymatic digests using previously published protocols.[239-241]

Double stranded DNA was obtained by combining equal amounts of the complementary strands (200 pmol) in 10 mM Tris-HCl (pH 8.0), 50 mM sodium chloride

buffer to achieve a concentration of 200 μM. The DNA mixtures were heated at 90°C for 5 mins and then allowed to cool slowly to room temperature. Each duplex contained a $^{15}N_3$,$^{13}C_1$-labeled guanine at the first position of codon 157, base paired to cytosine or a C5-alkylcytosine analog (**Table 2.2**).[242]

Table 2.1. Nucleobase sequences and molecular weights of oligodeoxynucleotides employed in this study.

Oligodeoxynucleotide ID	Sequence (5′ → 3′)	MW	
		Calc.	Obs.[#]
(+)*p53* exon 5	CCCGGCACCCGCG[*]TCCGCG	5715.7	5714.9
(-)*p53* exon 5	CGCGGA**C**GCGGGTGCCGGG	5911.9	5910.7
(-)*p53* exon 5-^{Me}C	CGCGGA**[Methyl-C]**GCGGGTGCCGGG	5925.9	5925.4
(-)*p53* exon 5-Et-C	CGCGGA**[Ethyl-C]**GCGGGTGCCGGG	5939.9	5939.5
(-)*p53* exon 5-prop-C	CGCGGA**[Propyl-C]**GCGGGTGCCGGG	5953.9	5954.2
(-)p53 exon 5-vinyl-C	CGCGGA**[Vinyl-C]**GCGGGTGCCGGG	5937.9	5939.0

[#] from HPLC-ESI-MS, G[*] = [$^{15}N_3$,$^{13}C_1$-G]

Table 2.2. Thermal melting points of structurally modified DNA duplexes employed in this work. Double-stranded DNA were obtained by dissolving equimolar amounts of the complimentary strands in 10 mM sodium phosphate buffer, pH 7.0 containing 50 mM NaCl to give a 9.7 μM final DNA concentration. DNA melting temperatures were obtained using a Varian Cary 100 Bio UV-visible spectrophotometer by linearly increasing the temperature by 0.5 °C/min from 30 to 90 °C. The melting points (T_m) was determined using Cary WinUV Thermal software (Varian, Palo Alto, CA) by averaging N = 3 runs. Data for native, methyl, ethyl, and propyl strands are reproduced from reference 9.[242]

Double stranded DNA	Sequence	T_m (ºC) [a]
ds native	5′-CCCGGCACCCGC G*TCCGCG-3′ 3′-GGGCCGTGGGCG CAGGCGC-5′	80.7 ± 0.5
ds methyl	5′-CCCGGCACCCGC G*TCCGCG-3′ 3′-GGGCCGTGGGCG ^{Me}CAGGCGC-5′	81.5 ± 0.2
ds ethyl	5′-CCCGGCACCCGC G*TCCGCG-3′ 3′-GGGCCGTGGGCG ^{Ethyl}CAGGCGC-5′	82.1 ± 0.2
ds propyl	5′-CCCGGCACCCGC G*TCCGCG-3′ 3′-GGGCCGTGGGCG ^{Propyl}CAGGCGC-5′	80.2 ± 1.1
ds vinyl	5′-CCCGGCACCCGC G*TCCGCG-3′ 3′-GGGCCGTGGGCG ^{Vinyl}CAGGCGC-5′	83.0 ± 0.8

[a] 9.7 μM dsDNA in 50 mM NaCl, 10 mM NaH_2PO_4, pH 7.0 (pH with NaOH). G* = [$^{15}N_3$,$^{13}C_1$-G]

DNA Methyltransferase Experiments

Synthetic DNA duplexes (250 -1500 fmol) were incubated with recombinant human DNMT1 enzyme (0.75U, New England Biolabs, Ipswich MA), 100 mg/ml BSA, and 160 μM S-adenosylmethionine (SAM) in 1X DNMT buffer for 15 min at 37 °C. The reactions were quenched by freezing on dry ice. The enzyme was inactivated by heating at 65 °C for 50 min. DNA was digested to 2′-deoxynucleosides in the presence of phosphodiesterase I (PDE I, 86 mU), phosphodiesterase II (PDE II, 77 mU), alkaline phosphatase (29 U), and DNase I (34 U) in 10 mM Tris-HCl/15 mM $MgCl_2$ buffer, pH 7 (35 μL total volume, 18 h at 37 °C). Samples were spiked with 2 pmol of $^{13}C_{10}{}^{15}N_2$-^{MeC} (internal standard for mass spectrometry) and purified by SPE using Thermo Fisher Scientific Hypersep Hypercarb (50 mg/mL) cartridges. Briefly, the cartridges were equilibrated with 2 mL MeOH and 2 mL deionized water. The volume of samples was adjusted to 500 μL prior to loading onto pre-equilibrated columns. The samples were washed with 2 mL each of deionized water and eluted in 1.6 mL of MeOH.

HPLC-ESI$^+$-MS/MS of 5-methyl-dC

The amounts of ^{MeC} present in DNA before and after incubation with DNMT1 were determined by HPLC-ESI$^+$-MS/MS utilizing a Thermo TSQ Vantage mass spectrometer interfaced with a Thermo Dionex Ultimate 3000. The mass spectrometer settings were optimized by direct infusion of standards at the same flow rate and composition of the eluting peaks. The mass source was typically operated at 2700 V, sheath gas 27,

declustering voltage 22 V, capillary temperature 270 °C, S-lens 93, collision gas 1.0 mTorr, and the collision energy at 8 V. The following MS/MS transitions were utilized: *m/z* 242.1 $[M + H]^+ \rightarrow$ *m/z* 126.1 [M – deoxyribose + H]$^+$ for ^{MeC} and *m/z* 254.1 $[M + H]^+ \rightarrow$ *m/z* 133.2 [M – deoxyribose + H]$^+$ for the corresponding internal standard $^{13}C_{10}{}^{15}N_2$-^{MeC}. Chromatographic separation was achieved using a Thermo Fisher Hypercarb column (0.3 mm x 100 mm, 3 μm) eluted at a flow rate of 14 μL/min at ambient temperature. HPLC solvents were 15 mM NH_4OAc (A) and ACN (B). A linear gradient of 15% to 60% B in 10 min was used, followed by an increase to 95% in 2 min and held at 95% B for 3 min. The column was returned to initial conditions over 2 min followed by equilibration at 15% B for 7 min. Under these conditions, both the ^{MeC} and $^{13}C_{10}{}^{15}N_2$-^{MeC} internal standard eluted at 6.1 min.

^{MeC} amounts introduced during DNMT1 reaction was calculated by comparing the HPLC-ESI-MS/MS peak areas corresponding to ^{MeC} and $^{13}C_{10}{}^{15}N_2$-^{MeC} internal standard. Values were adjusted by subtracting ^{MeC} amount generated spontaneously in the absence of the enzyme. The amount of DNMT1 mediated formation of ^{MeC} was used to calculate the velocity of methyl transfer. Plots of methylation velocities vs. DNA concentrations were generated for each DNA sequence (substrate) and fitted to a Michaelis-Menten equation below:

$$Y = V_{max}*X/ (K_m + X),$$

where the V_{max} is the maximal methylation rate of DNA by DNMT1, and K_m is the concentration of substrate required to reach the half-maximal rate. Statistical differences

in V_{max} were calculated in GraphPad Prism using ANOVA using Tukey correction for multiple testing.

Homology Modeling

All molecular modeling was performed using the Schrödinger modeling suite package.[243] Since both the mouse and human DNMT1 shared an 85% sequence similarity[244], homology modeling of the human DNMT1 (hDNMT1) was carried out using the published crystal structure of the mouse DNMT1 in complex with hemi-methylated DNA (PDB: 4DA4)[10] based on its reference sequence (NP_001124295.1). This method[245-246] takes advantage of the observation that protein structure is more highly conserved than its amino acid sequence, and that small or medium changes in sequence normally result in little variation in the 3D structure.[247] The DNA sequence was modified accordingly to match the sequence used experimentally to determine DNA methylation rates. Schrödinger Maestro (Schrödinger, LLC, NY) was used to model the extended forms of 5-methylcytosine (5-ethyl-dC, 5-propyl-dC, 5-vinyl-dC) within the modeled DNA template. Each of the modeled hDNMT1 – DNA complexes was subjected to standard protein preparation protocols at physiological pH, followed by energy minimization of the hydrogen atoms using OPLS3 force field with Generalized Born implicit solvent model to optimize all hydrogen-bonding networks.[248-249] To understand local structural effects induced by extension of 5-methylcytosine to $^{ethyl}C$, $^{propyl}C$, and $^{vinyl}C$, energy minimization of the modified base was carried out with a 15Å spherical radius restraint.

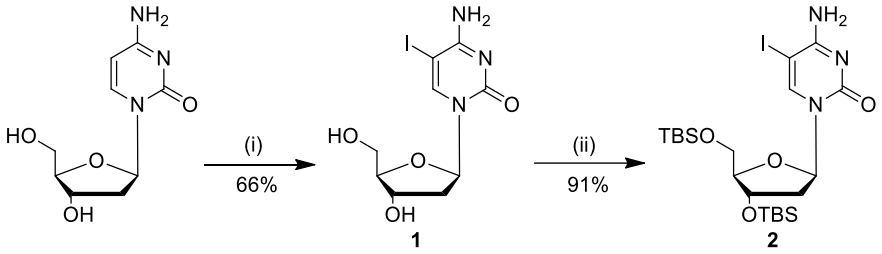
2.3 Results

2.3.1 Synthetic Strategy

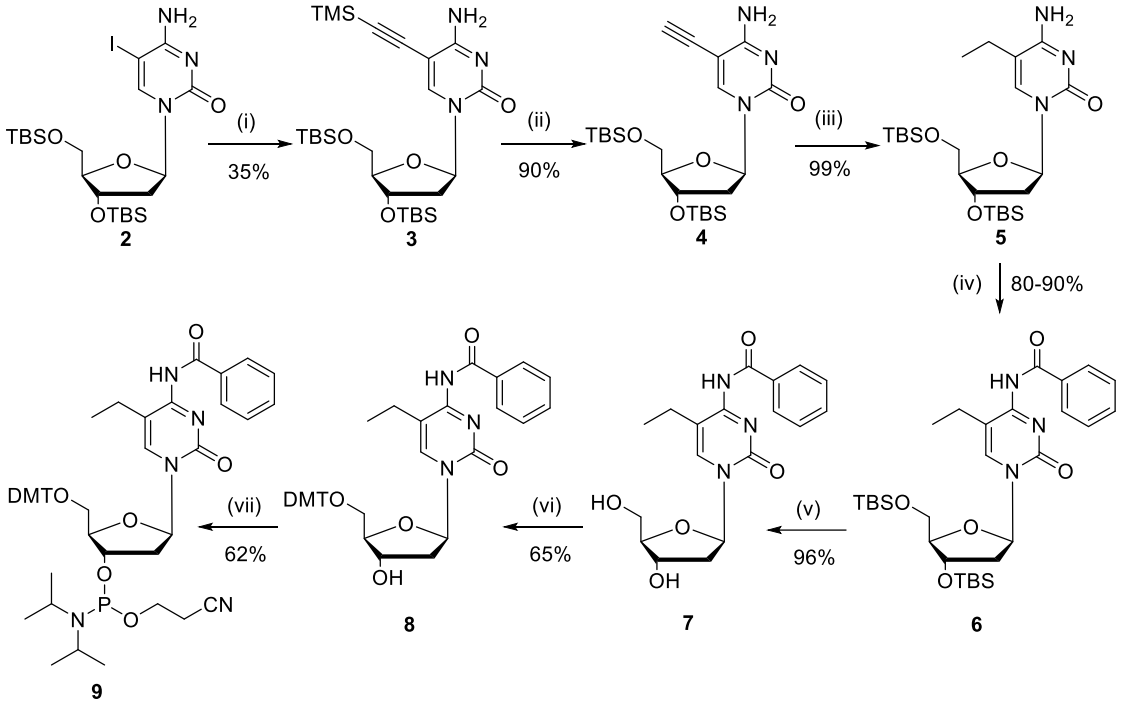
In order to incorporate 5-ethyl-dC, 5-propyl-dC, and 5-vinyl-dC into DNA strands, the corresponding nucleoside phosphoramidites were synthesized. The key starting material for the synthesis of these phosphoramidites was TBS-protected-5-iodo-dC (5-I-dC, **2** in **Scheme 2.2**).[250-251] 5-I-dC was subjected to Sonogashira and Stille cross- coupling reactions to introduce ethyl (**Scheme 2.3**), propyl (**Scheme 2.4**), or vinyl (**Scheme 2.5**) groups at the C-5 of dC.

C5-ethyl and propyl groups on cytidine were introduced by Sonogashira cross-coupling of compound **2** with TMS-protected acetylene (**Scheme 2.3**) or propyne gas (**Scheme 2.4**), respectively. The reactions were conducted in the presence of tetrakis(triphenyl-phosphine)palladium(0) at room temperature.[252-253] 5-(Trimethylsilyl)-ethynyl-3′,5′-O-*t*-butyldimethylsilyl-2′-deoxycytidine (**3**) and 5-propyn-1-yl-3′,5′-O-*t*-butyldimethylsilyl-2′-deoxycytidine (**10**) were obtained in 35% and 85% yield, respectively. The C-5 vinyl group was introduced via Stille coupling between **2** and tributylvinyltin in presence of palladium catalyst (**Scheme 2.5**). Compound **2** was reacted with tributylvinyltin in presence of palladium to get TBS-protected-5-vinyl-dC in 58% yield. The deprotection of the TMS group of compound **3** was achieved in the presence of potassium hydroxide in methanol to obtain compound **4** in 90% yield. The resulting 5-ethynyl-3′,5′-*t*-butyldimethylsilyl-2′-deoxycytidine **4** and compound **10** were subjected to hydrogenation to obtain the corresponding saturated compounds **5** and **11** in good yields.

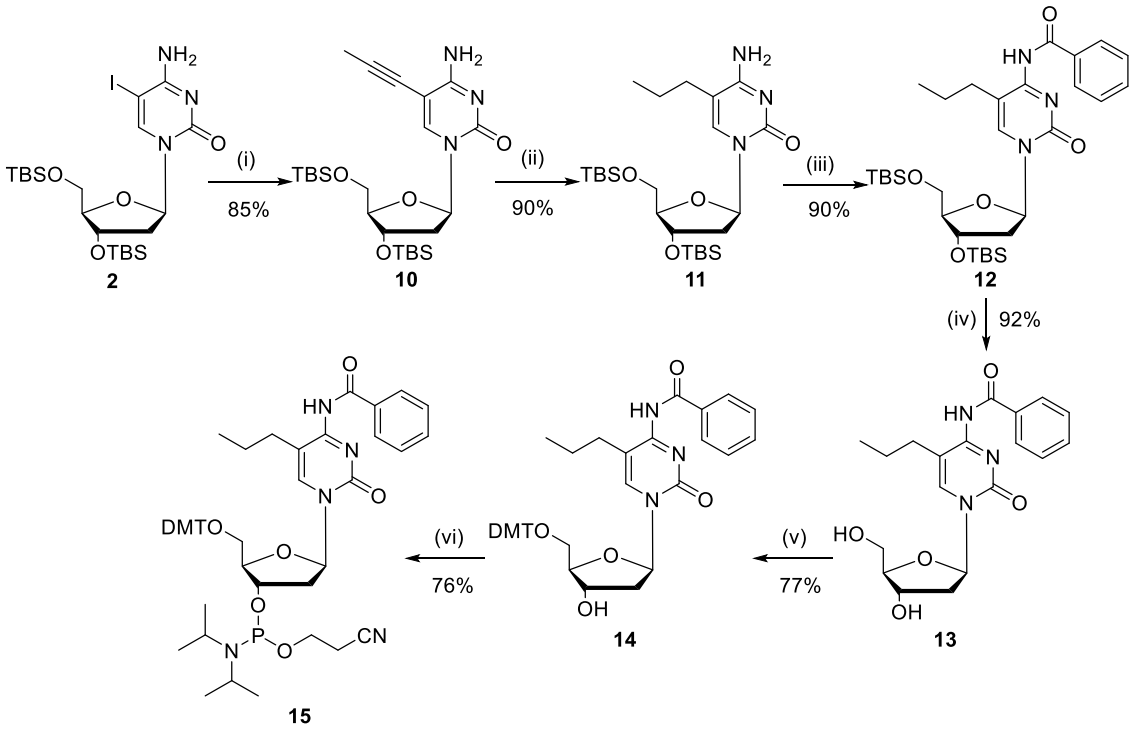
Scheme 2.2. Synthesis of common precursor 5-iodo-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**2**): (i) Iodine, iodic acid, water, glacial acetic acid, CCl_4 (ii) TBSCl, imidazole, DMF.



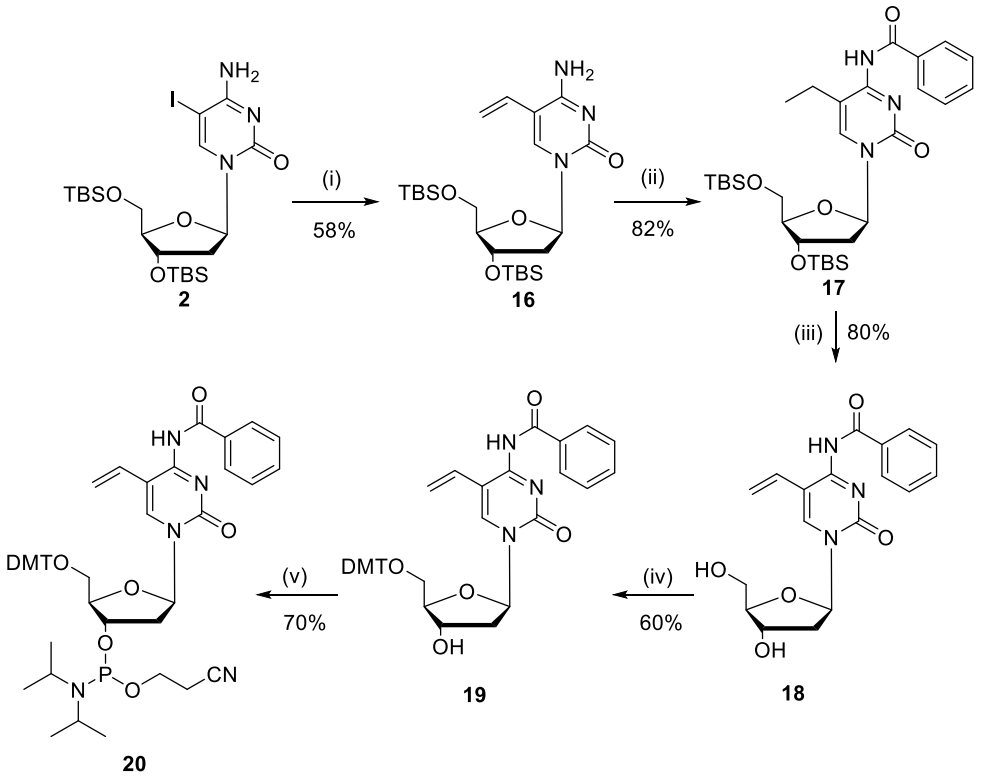
Scheme 2.3. Synthesis of 4-*N*-benzoyl-5′-DMT-5-ethyl-2′-dC-3′-phosphoramidite (**9**): (i) TMS acetylene, $Pd(PPh_3)_4$, CuI, anhydrous DMF, 48 h room temperature. (ii) 1 N KOH, anhydrous MeOH, 3 h room temperature. (iii) 10% Pd-C, H_2 (1atm), 24 h at room temperature. (iv) PhCOCl, anhydrous pyridine, overnight, room temperature. (v) Et_3N•3HF, anhydrous DCM, 16 h. (vi) DMTrCl, anhydrous pyridine, 16 h. (vii) 2-Cyanoethyl-*N,N*-diisopropylchlorophosphoramidite, DIPEA, anhydrous DCM, 16 h, room temperature.



Scheme 2.4. Synthesis of 4-*N*-Benzoyl-5′-DMT-5-propyl-2′-dC-3′-phosphoramidite (**15**): (i) propyne (5 psi), $Pd(PPh_3)_4$, CuI, anhydrous DMF, 24 h at room temperature. (ii) 10% Pd-C, anhydrous MeOH, H_2 (1atm), 24 h at room temperature. (iii) a. PhCOCl, anhydrous pyridine, overnight, room temperature. (iv) Et_3N•3HF, anhydrous DCM, 16 h. (v) DMTrCl, anhydrous pyridine, 16 h. (vi) 2-Cyanoethyl-*N,N*-diisopropylchlorophosphoramidite, DIPEA, anhydrous DCM, 16h, room temperature.



Scheme 2.5. Synthesis of 4-*N*-benzoyl-5′-DMT-5-vinyl-2′-dC-3′-phosphoramidite (**20**): (i) tributylvinyl tin, $Pd_2(dba)_3$, tri(2-furanyl)phosphine, anhydrous 1-methyl-2-pyrolidinone, 24 h 60 ℃ (ii) PhCOCl, anhydrous pyridine, overnight, room temperature. (iii**)** Et_3N•3HF, anhydrous DCM, 16 h. (iv) DMTrCl, anhydrous pyridine, 16 h. (v) 2-Cyanoethyl-*N*,*N*-diisopropylchloro phosphoramidite, DIPEA, anhydrous DCM, 16 h, room temperature.



TBS-protected 5-ethyl-dC, 5-propyl-dC, and 5-vinyl-dC nucleosides were converted to the corresponding phosphoramidites to enable their incorporation into DNA strands (**Schemes 2.3, 2.4, and 2.5**). In this process, nucleosides **5**, **11** and **16** were protected at exocyclic amine with benzoyl chloride to give the protected derivatives **6**, **12** and **17** in good yields (**6**, 80-90%; **12**, 90% and **17**, 82%). During the amine protection of the propyl-derivative (**13**) we observed a mixture of the expected mono-benzoylated (**12**) and the unexpected dibenzoylated product. This issue was resolved by stirring the crude product mixture in a mixture of 65:30:5 pyridine:MeOH:water containing 1N NaOH at 0 °C for 10 minutes and quenching the reaction with solid ammonium chloride.[234] Following benzoylation of the exocyclic amine, the TBS protecting groups were removed by treating the nucleosides **6**, **12** and **17** with triethylamine trihydrofluoride over 16 h in dry DCM to give nucleosides **7**, **13** and **18** in comparable yields (**Scheme 2.3, Scheme 2.4, and Scheme 2.5**). Further, the 5′-OH group of synthetic nucleosides was protected with DMT-Cl in anhydrous pyridine to give the corresponding DMT derivatives **8**, **14** and **19** in good to excellent yields. Phosphitylation of the DMT-ethers **8**, **14** and **19** in the presence of DIPEA with 2-cyanoethyl *N*,*N*-diisopropylchlorophosphoramidite resulted in the phosphoramidites 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-ethyl-2′-deoxycytidine-3′-[(2-cyanoethyl)-(*N,N*-diisopropyl)]-phosphoramidite (**9**) in **Scheme 2.3** and 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-propyl-2′-deoxycytidine-3′-[(2-cyanoethyl)-(*N,N*-diisopropyl)]-phosphoramidite (**15**) in **Scheme 2.4** and 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-vinyl-2′-deoxycytidine-3′-[(2-cyanoethyl)-(*N,N*-diisopropyl)]-phosphoramidite **20** in **Scheme 2.5**

[**9** (62%), **15** (76%), **20** (70%)]. All intermediates and final compounds were structurally characterized by a combination of ^{1}H-,^{13}C-,^{31}P-NMR, HRMS, and MS/MS (**Appendix A.1-A.19**).

The novel C5-functionalized dC phosphoramidites **9**, **15**, and **20** were used in solid phase synthesis to generate oligodeoxynucleotide 19-mers 5′-CGCGGA**[**alkyl-C**]**GCGGGT GCCGGG-3′ representing codons 155-160 of the p53 tumor suppressor gene (Table 2.1). In the resulting DNA duplexes (**Table 2.2**), structurally modified cytosine is base paired with guanine at the first position of *p53* codon 157, a prominent mutational "hotspot" in cancer.[254] All DNA strands were purified by reverse phase HPLC and characterized by ESI^{-}-MS (**Table 2.1** and **Figures 2.1-2.3**).[235-236] To obtain double stranded DNA, synthetic strands containing dC analogues were annealed to their compliments (**Table 2.2**). The duplexes were characterized by circular dichroism spectroscopy and thermal melting curves (**Figure 2.4** and **Table 2.2**).[236-238] These data indicated that the presence of the C5-alkylated cytidine analogs did not alter the overall secondary structure of the DNA duplexes, although small changes in T_m values were observed.

Figure 2.1. Deconvoluted ESI$^-$ mass spectrum of DNA oligomers: 5′-CGCGGA[**Et-C**]GCGGGTGCCGGG-3′ (M = 5939.9).

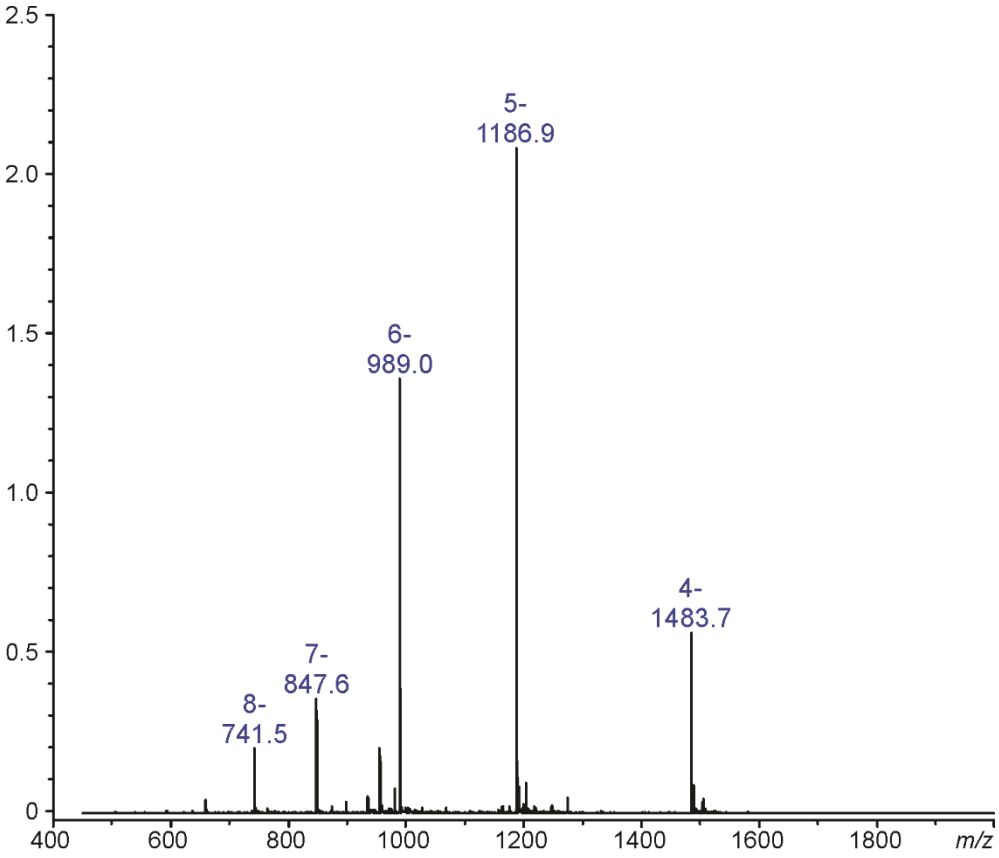


Figure 2.2. Deconvoluted ESI$^-$ mass spectrum of DNA oligomer: 5′-CGCGGA[**Pr-C**]GCGGGTGCCGGG-3′ (M = 5953.9).

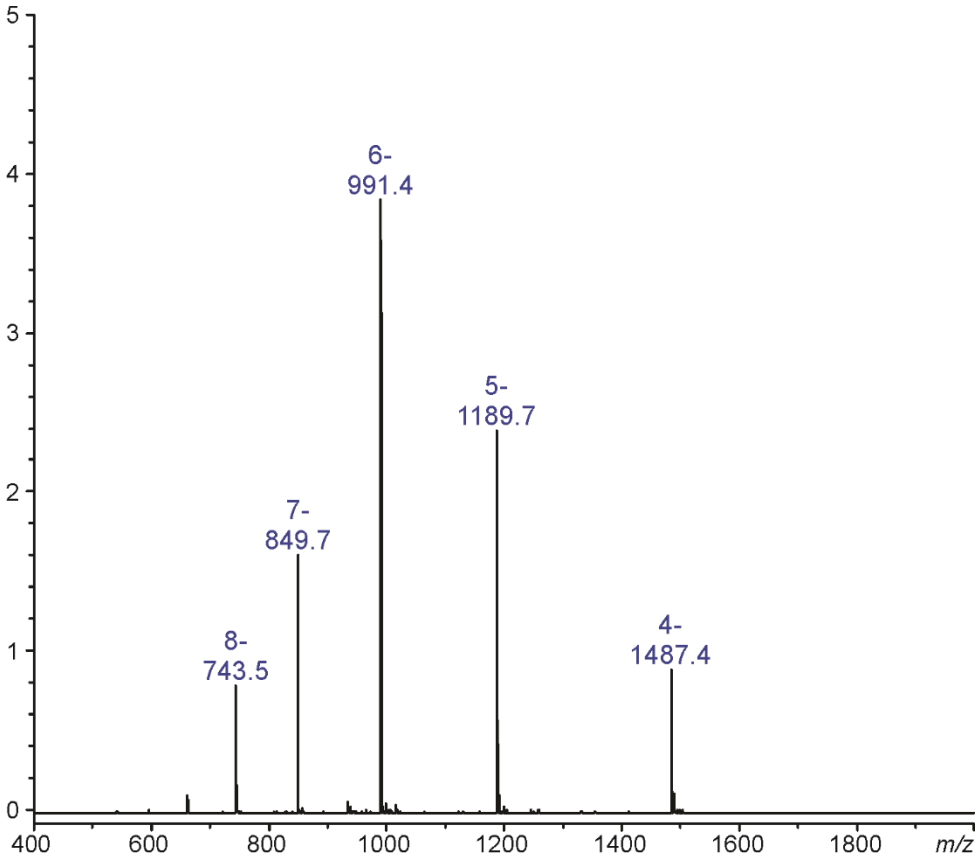


Figure 2.3. Deconvoluted ESI$^-$ mass spectrum of DNA oligomer: 5′-CGCGGA[**V-C**]GCGGGTGCCGGG-3′ (M = 5937.9).

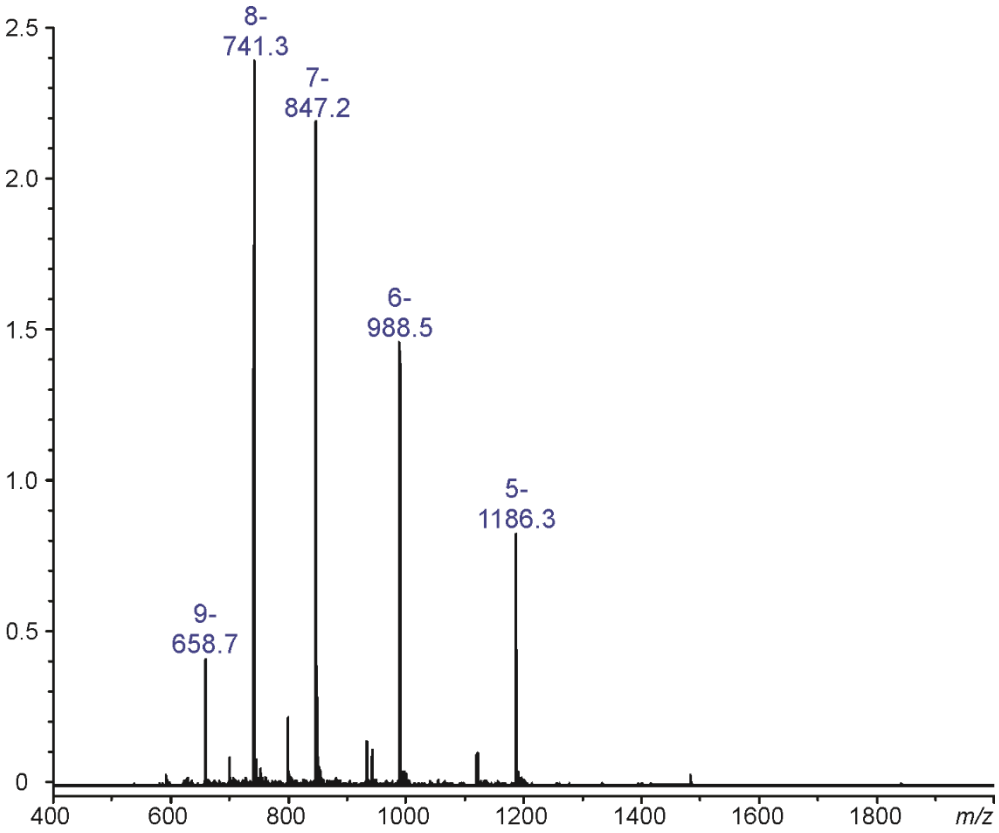
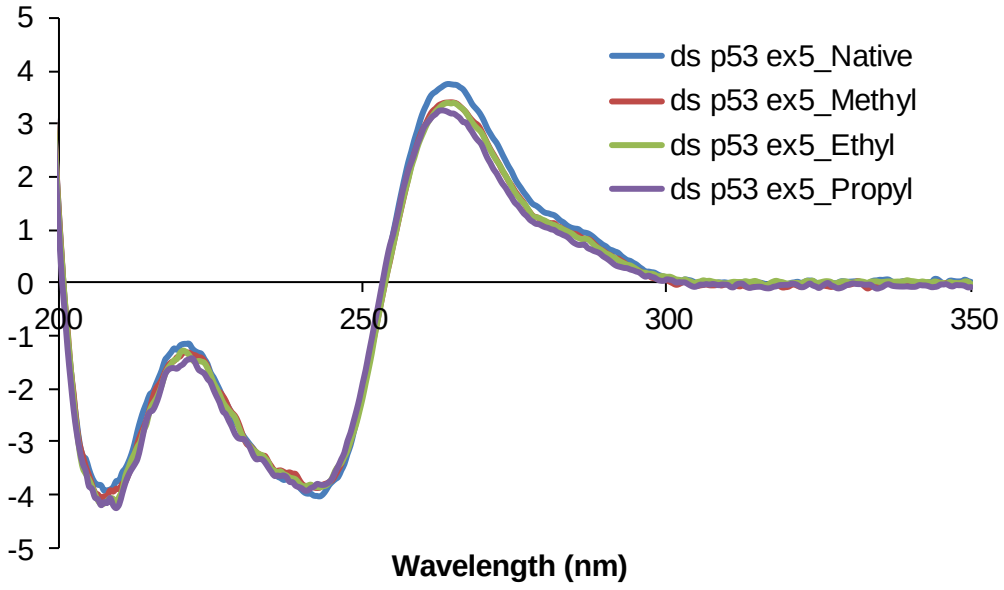


Figure 2.4. Circular Dichroism (CD) spectra of DNA oligonucleotides used in this work. Double-stranded DNA were obtained by dissolving equimolar amounts of the complimentary strands in 10 mM sodium phosphate buffer, pH 7.0 containing 50 mM NaCl to give a 9.7 μM final DNA concentration. The CD spectra were obtained with on a Jasco J-815 spectropolarimeter scanning the wavelengths from 350 to 200 nm in a 1 mm quartz cuvette, using a wavelength gradient of 0.5 nm. Spectra below are the average of N = 3 scans.



2.3.2 Influence of C-5 substituent on DNMT1-catalyzed cytosine methylation

To determine whether C-5 alkylcytosines with extended alkyl chain length are able to direct DNMT1 mediated cytosine methylation, DNA methylation assays were conducted with recombinant human DNMT1 and structurally modified duplexes containing 5-ethyl-dC, 5-propyl-dC, and 5-vinyl-dC in a CpG sequence opposite unsubstituted dC (**Scheme 2.6**). 5-methyl-dC was employed as a positive control, while unsubstituted dC served as a negative control for methylation reactions. Following incubation of DNMT1 and DNA duplexes in the presence of SAM and DNMT buffer, isotope dilution HPLC-ESI-MS/MS assay developed in our laboratory (**Scheme 2.6**) was employed to quantify the extent of DNMT1-mediated methyl transfer.

To quantify DNA methylation amounts, DNA was enzymatically digested to their corresponding 2′-deoxynucleosides and spiked with $^{13}C_{10}{}^{15}N_2$-^{MeC} internal standard, followed by solid phase extraction (SPE) (**Scheme 2.6**). ^{MeC} was quantified by capillary HPLC ESI$^+$-MS/MS using selected reaction monitoring of *m/z* 242.1 [M + H$^+$] → *m/z* 126.1 [M – deoxyribose + H$^+$] for MeC and *m/z* 254.1 [M + H$^+$] → *m/z* 133.2 [M – deoxyribose + H$^+$] for $^{13}C_{10}{}^{15}N_2$-^{MeC} internal standard (**Figure 2.5**). The amount of MeC formed was calculated by multiplying the amount of internal standard by the ratio of the areas for analyte and internal standard and subtracting the corresponding controls as shown in equation 1.

Equation 1: $MeC_{DNMT1} = \frac{Area_{analyte}}{Area_{ISTD}} * (ISTD) - MeC_{control}$

Scheme 2.6. Mass spectrometry based assay to quantify the rates of DNMT1-mediated methyl transfer in the presence of MeC and its analogues with extended C-5 side chain.

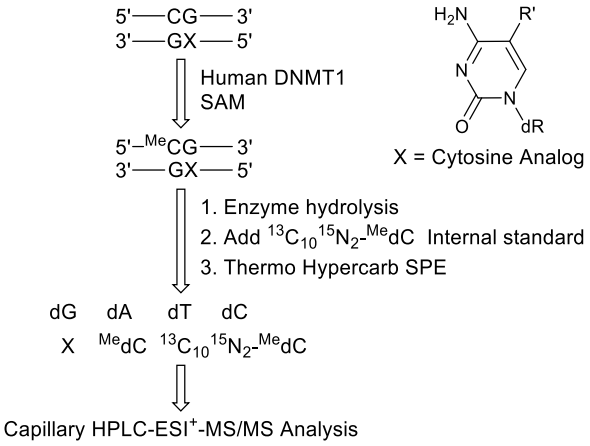
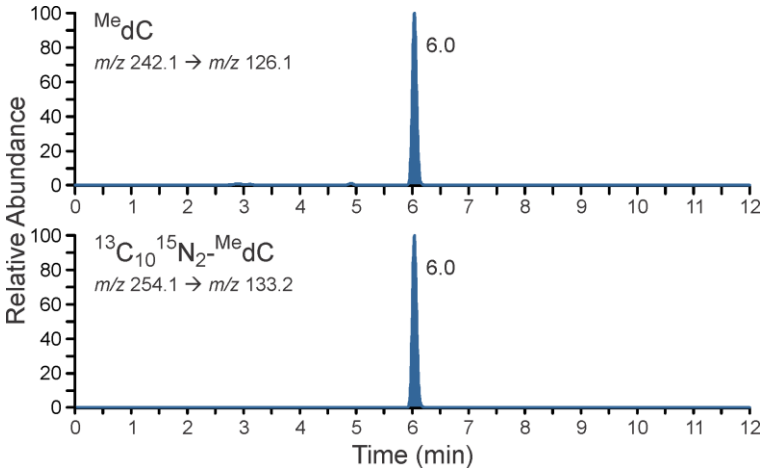


Figure 2.5. Capillary HPLC^{+}-ESI-MS/MS detection of 5-methyl-dC in methylation reactions.



Data were fitted into Michaelis-Menten curves to allow for kinetic analyses (**Figure 2.6**). The V_{max} and K_m values were determined via nonlinear regression using data from three or more individual points.[255] As anticipated, no DNMT1 mediated methylation was observed for DNA duplexes containing unmodified C:G base pair, while 5-methyl-dC containing DNA was efficiently methylated (**Figure 2.6, Table 2.3**). 5-ethyl-dC containing DNA was recognized as a DNMT1 substrate, although the V_{max} value was significantly reduced as compared to the native substrate (2.4×10^{-2} nM/min vs 9.6×10^{-2} nM/min, respectively, $p = 0.017$) (**Figure 2.6, Table 2.3**). The corresponding K_m values were 21.8 ± 13.7 and 3.90 ± 6.42 nM, for 5-methyl-dC and 5-ethyl-dC, respectively. In contrast, neither $^{Propyl}C$ nor $^{vinyl}C$ were able to direct DNA methylation under the conditions tested (**Figure 2.6**).

Figure 2.6. Michaelis-Menten plots showing activity of human DNMT1 toward DNA containing 5-alkyl-dC analogs (Sequence: 5′-CGCGGA[alkyl-C]GCGGGTGCCGGG-3′). The error bars represent the standard deviation of at least N = 3 repeats.

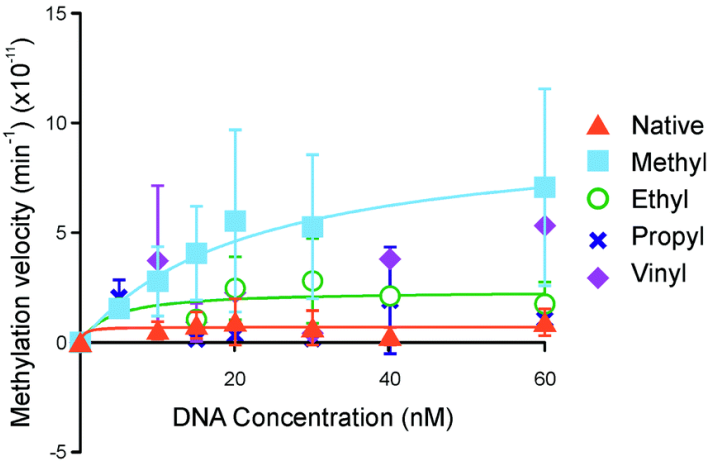


Table 2.3. Kinetic parameters of DNMT1 mediated methylation of DNA duplexes containing 5-alkyl-dC.[a]

DNA Duplex	V_{max} (x10^{-2} nM/min)	K_m (nM)	V_{max}/K_m (x10^{-2} min^{-1})
ds native	0.72 ± 0.22	0.450 ± 5.99	1.6 ± 21
ds methyl	9.6 ± 2.9	21.8 ± 13.7	0.44 ± 0.31
ds ethyl	2.4 ± 0.59	3.90 ± 6.42	0.62 ± 1.7

[a] The V_{max} and K_m values were determined via nonlinear regression using data from three or more individual points. The ranges in V_{max} and K_m are the standard error for regression analysis. Error was propagated for V_{max}/K_m using the equation: $\frac{Dc}{c} = \sqrt{\left(\frac{Da}{a}\right)^2 + \left(\frac{Db}{b}\right)^2}$; where a, b, and c are V_{max}, K_m, and V_{max}/K_m respectively.

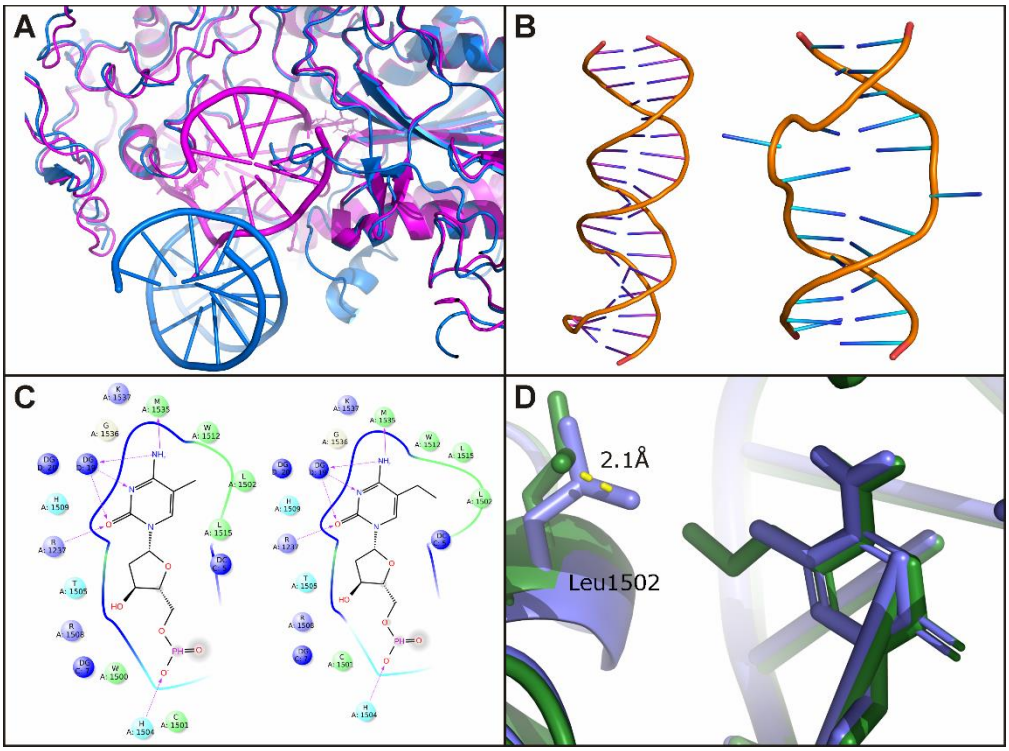
2.3.3 DNMT1 homology modeling

To determine whether the observed decrease in DNMT1 activity is associated with altered protein binding to structurally modified DNA, molecular modeling studies were conducted. Published crystal structures reveal two distinctive modes of binding between the unmethylated (PDB: 3PTA) and the hemi-methylated (PDB: 4DA4) DNA to DNMT1 (**Figure 2.7A**).[10, 244] In the catalytically productive complex, a large helical distortion around the central hemi-MeCpG site in the hemi-methylated DNA is observed (**Figure 2.7B**). To understand the effects of increasing C5-alkyl chain length on the activity of DNMT1, we generated a homology model of a productive hDNMT1 complex with hemi-methylated DNA.[245-246] A localized energy minimization encompassing all amino acid residues within 15 Å of the modified cytosine bases was performed to determine how the increased length of the alkyl chain influences the recognition of hemi-methylated CpG sites in DNA.

Within the productive hDNMT1-DNA complex, the target recognition domain (TRD) of the protein consists of a hydrophobic concave surface consisting of C1505, L1502, L1515, and M1535.[10] This pocket harbors the 5-methyl group of ^{Me}C and is involved in the recognition of the hemi-methylated ^{Me}CpG. We found that extension of the C5-alkyl chain disrupts DNA binding within the shallow hydrophobic pocket of the TRD due to the excessive steric bulk (**Figure 2.7C**). This distortion leads to a 2.1 Å displacement of L1502 in the case of 5-propyl-dC (**Figure 2.7D**) and 1.4 Å for 5-ethyl-dC. For the vinyl substitution, we also see a displacement of L1502 with the same magnitude of 1.4 Å as to

^{Ethyl}C. Thus, the difference of DNMT1 activity between 5-ethyl-dC and 5-vinyl-dC is likely due to the additional loss of rotational freedom by the co-planar geometry of the substitution.

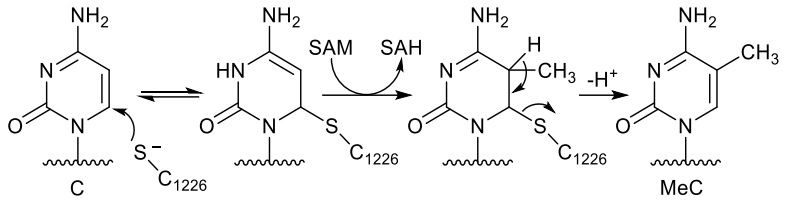
Figure 2.7. A. Crystal structures of human DNMT1 protein bound to methylated DNA (magenta)[10] or unmethylated DNA (blue)[244] reveal two distinctive binding orientations. **B.** Hemi-methylated DNA in the productive DNMT1 complex undergoes a large helical distortion. **C.** Excessive steric bulk from growing C5 alkyl chain distorts the ^{Me}C recognition pocket. **D.** ^{Propyl}C causes a 2.1 Å shift in the position of L1502.



2.4 Discussion

In cells, 5-methyl-dC at a hemimethylated CpG sites directs DNMT1 mediated DNA methylation of cytosine in the opposite strand. DNMT1 binds to hemimethylated DNA via the target recognition domain (TRD) of the protein, which is a hydrophobic pocket composed of M1535, C1501, L1502, L1515, and W1512.[10] Following recognition of ^{Me}C by the TRD, the cytidine to be methylated is flipped out of the duplex to enter the protein active site. The thiolate of the active site cysteine (C1226) forms a covalent bond with C-6, thereby activating the C5 position for methyl transfer from S-adenosylmethionine (SAM). Base catalyzed removal of the C5 proton allows for re-aromatization and release of the newly formed ^{Me}C (**Scheme 2.7**).[10, 256]

Scheme 2.7. Reaction mechanism of DNMT1 mediated methylation of cytosine.



In this work, we probed whether the TRD of DNMT1 could recognize C-5 alkylcytosine with an extended alkyl chain length and direct maintenance cytosine methylation using kinetic assays with human DNMT1. Using an isotope dilution, HPLC-ESI-MS/MS assay, we evaluated DNMT1 activity on synthetic DNA duplexes containing 5-ethyl-dC, 5-propyl-dC, and 5-vinyl-dC at a central CpG site opposite an unmodified dC. We found that DNMT1 was able to methylate 5-ethyl-dC containing DNA at a rate significantly less than the native substrate MeC (V_{max} = 2.4 x 10^{-2} nM/min vs. 9.6 x 10^{-2} nM/min, respectively), but there was no methyl transfer activity for DNA containing 5-vinyl-dC or 5-propyl-dC (**Figure 2.6 and Table 2.3**).

Our data suggests that the increase in steric bulk at C-5 by extending the alkyl chain length hinders DNMT1 enzyme activity. Although DNMT1 is able to recognize and methylate the DNA sequence containing 5-ethyl-dC, duplexes containing 5-propyl-dC are poor DNMT1 substrates (**Figure 2.6**). Valinluck et al. previously reported that DNMT1 was capable of accommodating DNA duplexes containing 5-chloro-dC, 5-bromo-dC, and iodo-dC, but not the more rigid 5-propynyl-dC.[231] In contrast, oxidation of MeC to 5-hydroxymethyl-dC, 5-formyl-dC, and 5-carboxyl-dC hindered DNMT1 activity.[233] In addition, 5-vinyl-dC containing duplex was not a substrate for DNMT1 (**Figure 2.6**).

Upon observing the reduced activity of DNMT1 toward DNA containing 5-ethyl-dC and inactivity on DNA containing 5-vinyl-dC or 5-propyl-dC, we explored the potential for structural changes in the MeC recognition site when DNA contained ^{Propyl}C, ^{Ethyl}C, and ^{Vinyl}C using molecular modeling. A homology model for hDNMT1 was created by

replacing the sequence of mouse DNMT1 in the published crystal structure of the catalytically productive complex (PDB: 4DA4) with the sequence human DNMT1.[10] Using localized energy minimization around the modified cytosine, we found that modification of C5-alkyl substituent displaces the residues within the TRD, especially L1502 (**Figures 2.7C** and **2.7D**). The magnitude of displacement of the nearby L1502 residue is largest for the bulkiest alkyl group propyl at 2.1 Å and less significant for ethyl and vinyl 1.4 Å (**Figure 2.7D** and **Table 2.3**). The large disruptions needed in the TRD to accommodate extended alkyl side chains reduce the recognition of C5-alkyl cytosines by DNMT1 and prevent the formation of a productive DNA-DNMT1 complex.

Our model supports the kinetic data shown in **Figure 2.6** and **Table 2.3**, with the K_m values reflecting either a productive or unproductive mode of binding. For unmethylated DNA, we observe the lowest K_m value (0.450 nM, Table 2.3), demonstrating that in the unproductive DNMT1-DNA complex, DNA is more tightly bound. The productive DNMT1-DNA complex is hallmarked by a large distortion in the DNA duplex and a change in the mode of binding, resulting in a lower K_m value.[10, 244] As the length of the C5-alkyl chain increases, it causes a disruption in the TRD pocket. This hinders the recognition of the hemi-methylated MeCpG, causing the enzyme to adopt an unproductive mode of tighter binding with a lower K_m value.

In summary, we have synthesized DNA duplexes containing a range of 5-alkyl-dC analogues with extended side chain. These unnatural DNA substrates were used to explore the ability of DNMT1 protein to recognize and methylate CpG sequences containing C5-

alkylcytosine modifications with extended alkyl chain. The enzyme was capable of recognizing and methylating DNA containing ^{Me}C and $^{Ethyl}C$ with similar efficiencies, indicating that the ethyl group can be tolerated. In contrast, DNMT1 showed no enzymatic activity towards $^{Vinyl}C$ and $^{Propyl}C$, containing duplexes under the conditions tested (**Figure 2.6** and **Table 2.3**). We hypothesize that DNMT1 can bind to DNA to form two distinct complexes: a tighter binding inactive conformation and a lower affinity active conformation. Moreover, molecular models of DNMT1-DNA interactions indicate that the presence of larger C-5 alkyl chains at dC disrupt the integrity of the DNA recognition pocket, with $^{Propyl}C$ pushing DNMT1 conformation toward the inactive complex. Although $^{Ethyl}C$ and $^{Vinyl}C$ cause a smaller disruption of the DNA recognition pocket, only $^{Ethyl}C$ duplex is a substrate for DNMT1, suggesting that the presence of a C-5 vinyl group affects other steps within the catalytic cycle.

III. MAINTENANCE DNA METHYLTRANSFERASE ACTIVITY IN THE PRESENCE OF OXIDIZED FORMS OF 5-METHYLCYTOSINE: STRUCTURAL BASIS FOR TET-MEDIATED DNA DEMETHYLATION

Adapted from:

Seiler, C.L.; Fernandez, J.; Koerperich, Z.; Andersen, M.P.; Kotandeniya, D.; Nguyen, M.E.; Sham, Y.Y.; and Tretyakova N.Y.; Maintenance DNA methyltransferase activity in the presence of oxidized forms of 5-methylcytosine: Structural basis for TET-mediated, DNA demethylation, *Biochemistry*, Accepted with revision August 2018.

This work was performed in collaboration with Jenna Fernandez, Zoe Koerperich, Molly P. Andersen, Dr. Delshanee Kotandeniya, Megin E. Nguyen, Dr. Yuk Y. Sham, under the direction of Dr. Natalia Y. Tretyakova. Christopher Seiler performed the kinetic assays with the assistance of Molly Andersen. Christopher Seiler and Molly Andersen synthesized, and purified DNA strands used in this work. Jenna Fernandez performed the homology modeling and molecular dynamics simulations with the assistance of Megan Nguyen under the guidance of Dr. Sham. Christopher Seiler and Zoe Koerperich performed the electrophoretic mobility shift assays. Dr. Kotandeniya developed the kinetic assay of DNMT1 activity.

3.1 Introduction

5-Methylcytosine (MeC) is a stable epigenetic mark most commonly found at CpG dinucleotides of DNA, where cytosine bases in both strands are methylated.[257] Cytosine methylation typically has a repressive effect, leading to reduced levels of gene expression.[9-10] Methylated CpG sites within gene promoter regions interfere with transcription factor binding and instead are recognized by methyl-CpG binding proteins, promoting the recruitment of histone deacetylases and the formation of closed chromatin.[9-10]

In human cells, DNA methylation marks are introduced by *de novo* methyltransferases DNMT3a/b and are subsequently preserved by maintenance DNA methyltransferase 1 (DNMT1).[12] The activity of DNMT1 as a maintenance methyltransferase during replication is necessary to ensure accurate transmission of epigenetic methylation marks to progeny cells. DNMT1 is recruited to the replication fork by UHRF1 and acts at hemimethylated CpG sites generated during DNA replication through recognition of the methylated CG sites on the template strand of the DNA.[258-259] DNMT1 specifically recognizes and methylates hemi-methylated CG sites.[10, 244] Formation of an enzymatically active complex requires MeC binding to a concave hydrophobic pocket within the target recognition domain (TRD) of DNMT1.[10, 260] This induces local melting of the DNA duplex, allowing unmethylated cytosine in the opposite strand of DNA to be actively "flipped" out of the DNA duplex stack to enter the protein active site.[10, 258, 260-261] The reversible addition of a thiolate from an active site cysteine residue of DNMT1 to the C-6 position of cytosine activates the C-5 position of the nucleobase, allowing it to accept

a methyl group from the S-adenosylmethionine (SAM) cofactor.[10, 262] The covalent DNA-protein complex is reversed via base catalyzed removal of the H-5 proton of the nucleobase and re-aromatization via elimination of the covalently attached cysteine (**Scheme 2.7 in Chapter II**).[10]

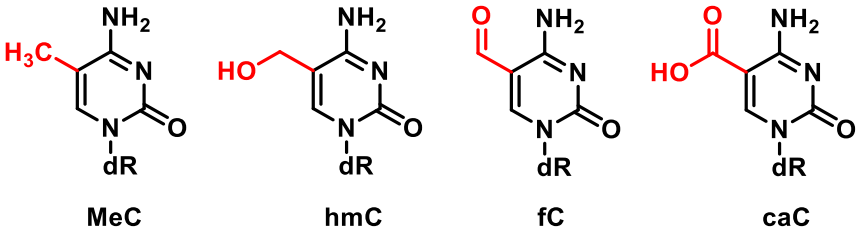
Crystal structures reveal two distinctive modes of DNMT1-DNA binding.[10, 244] In productive DNMT1-DNA complex containing a hemi-methylated CG dinucleotide (PDB:4DA4), hydrophobic amino acid side chains of M1535, C1501, L1502 and L1515 form a target recognition domain (TRD) involved in interactions with hemimethylated MeCpG.[10] As a hemimethylated CpG site emerges from the replication complex, the TRD of DNMT1 specifically recognizes MeC and facilitates maintenance methylation. A different binding mode is revealed in the crystal structure of DNMT1 interacting with unmethylated DNA (PDB:3PTA).[244] In the unproductive DNMT1-DNA complex, the zinc finger-containing CXXC domain of the protein prevents *de novo* methylation by positioning the CXXC-BAH1 linker region of the protein between the DNA and the catalytic active site.[244] Furthermore, the TRD of the protein assumes a retracted position that prevents it from direct interaction with the DNA.[244] This autoinhibitory mechanism protects DNA from unintended *de novo* methylation by DNMT1.[244]

Ten Eleven Translocation (TET) dioxygenases sequentially oxidize the methyl group of MeC in DNA to give 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) (**Scheme 3.1**).[13, 15-16] Both fC and caC can be excised by thymine DNA glycosylase (TDG) and replaced with cytosine via TDG-mediated base

excision repair pathway, leading to active demethylation.[263] Additionally, oxidized forms of MeC promote passive DNA demethylation by interfering with DNMT1 activity at hemimethylated CpG sites.[19] The relative contributions of passive and active DNA demethylation are dependent on the cell type, developmental factors, and the stage of the cell cycle.[264-265] However, to our knowledge, the kinetics of DNMT1-mediated methyl transfer in the presence of oxidized forms of MeC (oxo-MeC) has not been elucidated, and the structural origins of reduced activity of DNMT1 protein in the presence of oxo-MeC are not well understood.

In the present study, a mass spectrometry based quantitative assay developed in our laboratory was used to examine the kinetics of DNMT1-mediated maintenance methylation in the presence of MeCG, hmCG, fCG, and caCG, while molecular dynamics simulations were conducted to examine the structural origins of the reduced DNMT1 activity in the presence of oxidized forms of 5-methylcytosine.[20] Our results support a model in which MeC oxidation to hmC, fC, and caC prevents the formation of a productive DNMT1- DNA complex by weakening hydrophobic interactions between the modified cytosine and the TRD of the DNMT1 protein, leading to reduced maintenance methylation rates and allowing for passive DNA demethylation.

Scheme 3.1: Epigenetic modifications of cytosine in DNA.



3.2 Materials and Methods

Materials

All nucleoside phosphoramidites including 5-methyl-dC, 5-hydroxymethyl-dC, 5-formyl-dC-III, 5-carboxy-dC, Ac-dC, dT, dA, dG, and dmf-dG, reagents, and controlled pore glass solid support for oligodeoxynucleotide synthesis were acquired from Glen Research Corporation (Sterling, VA). Human recombinant DNA methyltransferase 1 (DNMT1) and Δ580-DNMT1[266] (missing the PCNA,[267] DNMT3A/B interaction domains[267-268]) were purchased from New England BioLabs (Ipswich, MA). Phosphodiesterase I, phosphodiesterase II, and DNase I were acquired from Worthington Biochemical Corporation (Lakewood, NJ). Bovine intestinal alkaline phosphatase was procured from Sigma Aldrich Chemical Company (Milwaukee, WI). All remaining laboratory chemicals and solvents were purchased from ThermoFisher Scientific (Waltham, MA) and Sigma-Aldrich (Milwaukee, WI). The synthesis of $^{13}C_{10}{}^{15}N_2$-5-Methyl-2′-deoxycytidine was described previously.[20]

Synthesis of MeC, hmC, fC and caC Containing Oligodeoxynucleotides

Synthetic DNA oligodeoxynucleotides (**Table 3.1**) were assembled by solid phase synthesis on an ABI 394 DNA synthesizer (Applied Biosystems, Grand Island, NY). MeC and its oxidized forms (MeC, hmC, fC, and caC) were added via manual coupling.[242] MeC containing strands were cleaved from solid support and deprotected using 30% ammonium hydroxide for 16 h at room temperature. hmC containing DNA was cleaved from support

and deprotected using 30% ammonium hydroxide at 75 °C for 16 h. fC containing DNA was cleaved from solid support by incubation in 30% ammonium hydroxide for 16 h, followed by desalting using Illustra® Nap-5 cartridges (GE Healthcare, Buckinghamshire, UK). The 5-(1,3-dioxane-2-yl) protecting group on fC was cleaved using glacial acetic acid for 6 h at room temperature to reveal the unprotected oligodeoxynucleotide. Synthetic DNA strands containing caC were cleaved and deprotected using 0.4 M methanolic sodium hydroxide (80% methanol:20% 2 M sodium hydroxide) overnight. Following the removal of the supernatant and neutralization of the base with 2 M triethyl ammonium acetate, silica solid support was sonicated with water 3 times to maximize DNA yields.

Table 3.1: Nucleotide sequences and mass spectrometry characterization (HPLC-ESI$^-$-MS/MS) of DNA oligomers used in methylation experiments. Measurements were taken on Agilent MSD Ion Trap operating in the negative ion mode.

Sequence (5′-3′)	Calculated MW	Observed MW
TCAGATTCGCGC**C**GGCTGCGATAAGCT	8276.4	8275.9
AGCTTATCGCAGC**C**GGCGCGAATCTGA	8285.4	8285.1
AGCTTATCGCAGC^{Me}**CGG**CGCGAATCTGA	8299.5	8299.4
AGCTTATCGCAGC^{hm}**C**GGCGCGAATCTGA	8315.5	8315.0
AGCTTATCGCAGC^{f}**C**GGCGCGAATCTGA	8313.4	8313.5
AGCTTATCGCAGC^{ca}**C**GGCGCGAATCTGA	8329.4	8329.0
AGCTTATCGCAGC^{Me}**CGT**CGCGAATCTGA	8274.5	8274.4
AGCTTATCGCAGC^{Me}**CGA**CGCGAATCTGA	8283.5	8283.4
AGCTTATCGCAGC^{Me}**CGC**CGCGAATCTGA	8259.4	8259.4
TCAGATTCGCG**ACG**GCTGCGATAAGCT	8300.5	8300.4
TCAGATTCGCG**TCG**GCTGCGATAAGCT	8291.4	8291.4
TCAGATTCGCG**GCG**GCTGCGATAAGCT	8316.5	8316.5

Synthetic DNA strands were purified by reverse phase HPLC using an Agilent 1100 HPLC system interfaced with a UV variable wavelength detector set at 260 nm. A Varian Pursuit C-18 HPLC column (5 µm, 250 x 10.0 mm) was eluted at flow rate of 3 mL/min with a gradient of 100 mM triethylammonium acetate pH 7.0 (A) and acetonitrile (B). In method A, solvent composition was linearly changed from 8.4% to 12% B over 30 min, increased to 17.5% over 10 min and further to 32.5% over 5 min. Solvent composition returned to initial conditions over 2 min, followed by re-equilibration for 15 min. In method B, solvent composition was changed from 9% to 16.3% B over 30 min, increased to 20% over 10 min, further to 32.5% over 5 min, and returned to initial conditions over 2 min, followed by equilibration over 15 min. Method A was used to purify C, MeC, hmC, and caC containing strands while method B was used to purify fC containing DNA.

Following HPLC purification, synthetic DNA strands were desalted using Illustra Nap-5 cartridges (GE Healthcare) according to the manufacturer's instructions. The presence of MeC, hmC, fC, or caC in synthetic DNA strands was confirmed by HPLC-ESI$^-$-MS analyses on an Agilent MSD Ion Trap interfaced with an Agilent 1100 HPLC system (**Table 3.1**). A Zorbax 300-SB C-18 column (5 µm, 150 x 0.5 mm) was eluted with a gradient of 15 mM ammonium acetate (A) and acetonitrile (B). Solvent composition was held at 2% B for 3 min, then linearly increased to 40% over 16 min, further to 55% over 1 min, followed by a return to initial conditions over 1 min and equilibration for 13 min. DNA concentrations were determined by quantitation of the 2′-deoxyguanosine in enzymatic digests.[239-241]

DNA Methyltransferase experiments

Human DNMT1 (0.75U, New England BioLabs, Ipswich, MA) was incubated with synthetic DNA duplexes (250 – 1500 fmol) containing a single, centrally located CpG site with MeC, hmC, fC, or caC opposite the target C (**Table 3.2**). Methylation reactions were conducted in commercial DNMT buffer in the presence of 0.1 mg/mL bovine serum albumin (BSA) and 160 μM S-adenosylmethionine (SAM) for 15 min at 37 °C. The reactions were quenched by placing samples on dry ice, followed by enzyme inactivation by heating at 65 °C for 40 min. DNA was digested to 2′-deoxynucleotides in the presence of PDE I (55 mU), PDE II (63 mU), DNase I (28 U), and alkaline phosphatase (48 U) in a solution containing 10 mM Tris-HCl pH 7.0 and 15 mM magnesium chloride. Samples were spiked with $^{13}C_{10}{}^{15}N_2$-5-methyl-2′-deoxycytidine (internal standard for mass spectrometry, 1.33 pmol) and purified by offline HPLC. A Waters Atlantis T3 column (3 μm, 4.6 mm x 150 mm) was eluted with a gradient of 5 mM ammonium formate pH 4.0 (A) and methanol (B). Solvent composition was linearly changed from 3% B to 20% B in 15 min, further to 40% B over 5 min, increased to 80% over 5 min, held at 80% for 2 min, and returned to initial conditions over 2 min, followed by column reequilibration for 8 min. HPLC fractions containing MeC and its internal standard (8.9 – 10.2 min) were collected, concentrated under reduced pressure, and reconstituted in 12 μL of 15 mM ammonium acetate buffer prior to HPLC-ESI$^+$-MS/MS analysis.

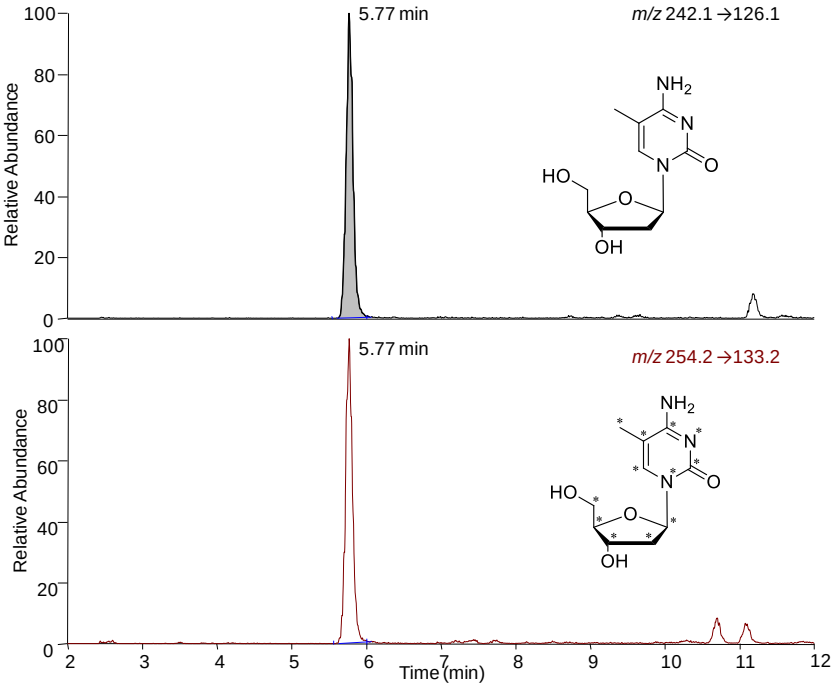
Table 3.2: DNA duplexes used to investigate the effects of TET oxidation and local sequence on the rate of cytosine methylation mediated by DNMT1.

Duplex Name	Duplex sequence	Melting temperature $T_m \pm SD$ (°C)
(-)C	5′-TCAGATTCGCGCCGGCTGCGATAAGCT-3′ 3′-AGTCTAAGCGCGG**C**CGACGCTATTCGA-5′	77.4 ± 0.2
(-)MeC	5′-TCAGATTCGCGCC GGCTGCGATAAGCT-3′ 3′-AGTCTAAGCGCGG^{Me}**C**CGACGCTATTCGA-5′	77.0 ± 0.6
(-)hmC	5′-TCAGATTCGCGCC GGCTGCGATAAGCT-3′ 3′-AGTCTAAGCGCGG^{hm}**C**CGACGCTATTCGA-5′	76.6 ± 0.9
(-)fC	5′-TCAGATTCGCGCC GGCTGCGATAAGCT-3′ 3′-AGTCTAAGCGCGG^{f}**C**CGACGCTATTCGA-5′	77.5 ± 0.6
(-)caC	5′-TCAGATTCGCGCC GGCTGCGATAAGCT-3′ 3′-AGTCTAAGCGCGG^{ca}**C**CGACGCTATTCGA-5′	78.0 ± 0.4
MeCGT	5′-TCAGATTCGCG**A**C **G**GCTGCGATAAGCT-3′ 3′-AGTCTAAGCGC**T**G^{Me}**C**CGACGCTATTCGA-5′	75.7 ± 0.3
MeCGA	5′-TCAGATTCGCG**T**C **G**GCTGCGATAAGCT-3′ 3′-AGTCTAAGCGC**A**G^{Me}**C**CGACGCTATTCGA-5′	75.5 ± 0.7
MeCGC	5′-TCAGATTCGCG**G**C **G**GCTGCGATAAGCT-3′ 3′-AGTCTAAGCGC**C**G^{Me}**C**CGACGCTATTCGA-5′	77.3 ± 0.4

HPLC-ESI$^+$-MS/MS analysis

Quantification of MeC was carried out by isotope dilution HPLC-ESI$^+$-MS/MS with $^{13}C_{10}{}^{15}N_2$-MeC as an internal standard. A Thermo Dionex Ultimate3000 HPLC system was coupled to a Thermo TSQ Vantage mass spectrometer (ThermoFisher Scientific, Waltham, MA). A Thermo Hypercarb column (3 μm, 100 x 0.5 mm) was maintained at 60 °C and eluted with a gradient of 15 mM ammonium acetate (A) and acetonitrile (B). Solvent composition was changed from 15 to 60% B over 10 min, further to 95% over 2 min, held at 95% B for 3 min, and returned to initial conditions over 2 min, followed by re-equilibration for 7 min. HPLC eluent was directed into the mass spectrometer during 2-12 min of the chromatographic run. Typical MS parameters were as follows: spray voltage, 3200 V; sheath gas pressure, 20 psi; capillary temperature 350 °C; collision energy, 8; declustering voltage, 22 V; collision gas pressure, 1.5 mTorr; tuned S-Lens, 93; Q1 (full width at half maximum), 0.4; Q3 (full width at half maximum), 0.7; scan width, 0.4; scan time, 0.1 s. The mass spectrometer parameters were optimized upon direct infusion of authentic standards of MeC. The instrument was operated in the selected reaction monitoring mode by following the transitions *m/z* 254.1 $[M + H]^+ \rightarrow$126.1 $[M + H - dR]^+$ for MeC and *m/z* 254.1$[M + H]^+ \rightarrow$133.1 $[M + H - dR]^+$ for $^{13}C_{10}{}^{15}N_2$-MeC (**Figure 3.1**). MeC amounts in each sample were determined from HPLC-ESI$^+$-MS/MS peak areas corresponding to the analyte and its internal standard using calibration curves constructed with authentic standards.

Figure 3.1: HPLC-ESI^{+}-MS/MS analysis of MeC and $^{13}C_{10}{}^{15}N_2$-MeC in hydrolysates of DNA following DNMT1 treatment and offline-HPLC fractionation.



Methylation velocity (V_{MeC}, M/min) was calculated from the HPLC-ESI^+-MS/MS areas corresponding to the analyte and internal standard according to the following equation:

$$V_{MeC} = (A_{AN} / A_{IS}) * C_{IS} / [(V * t) * 1 \times 10^{-6}],$$

where A_{AN} and A_{IS} are the areas under the HPLC-ESI^+-MS/MS peaks corresponding to MeC (analyte, AN) and its ^{13}C, ^{15}N-labeled internal standard (IS), respectively, C_{IS} is the amount of internal standard used in pmol, V is the volume in microliters, and t is the reaction time in min. Steady-state kinetic parameters (K_m and V_{max}) for methyl transfer reaction were determined by plotting the calculated velocities vs. substrate concentration.

Kinetic Analyses

The MeC amounts determined by HPLC-ESI^+-MS/MS were plotted against DNA concentrations using Prism 6 software from GraphPad Software, Inc. (La Jolla, CA). The kinetic curves were fitted to the Michaelis-Menton equation using non-linear regression to give the values of K_m and V_{max}.

Electrophoretic Mobility Shift Assay (EMSA) to study DNMT1-DNA binding

DNA strands (**Table 3.1**) (50 pmol) were radiolabeled by incubation with T4 polynucleotide kinase (20 U, New England BioLabs, Beverly, MA) and γ^{32}P-ATP (5 µCi, PerkinElmer Life Sciences, Boston, MA) in 1X polynucleotide kinase reaction buffer for 1 hour at 37 °C. The enzyme was inactivated by heating at 65 °C for 10 min, and the free

γ^{32}P-ATP was removed by Illustra MicroSpin G-25 Column (GE Healthcare, Pittsburgh, PA). After radiolabeling the forward and reverse strands, the complementary DNA strands were annealed by heating to 90 °C for 5 min followed by slowly cooling to room temperature to obtain double stranded DNA substrates (**Table 3.2**).

^{32}P-end-labeled DNA duplexes (**Table 3.2**) containing MeC, hmC, fC, or caC (2 nM) were incubated with purified human Δ580-Dmnt1 (0-128 nM, New England BioLabs, Beverly, MA), and 1X gel shift assay buffer (10 mM HEPES, 50 mM KCl, 0.1 mM EDTA, 1 mM DTT 2.5 mM $MgCl_2$, 0.2% Triton X-100, and 10% glycerol) at 37 °C for 30 min. The mixture was then loaded onto 4 % polyacrylamide gel (37.5:1 acrylamide:bisacrylamide ratio, prepared with 0.5X TBE) while running at 300 V for 10 minutes. The gels were electrophoresed at 140 V for an additional 40 minutes at 4 °C. Gels were imaged with a Typhoon FLA 7000 instrument (GE Healthcare).

Homology Modeling

All molecular modeling simulations were performed using the Schrödinger modeling suite package (Schrödinger, LLC, NY).[243] Homology modeling was carried out using Schrodinger's Prime as described previously.[20] In brief, homology modeling of the human DNMT1 (hDNMT1) utilized the crystal structure of mouse DNMT1 (mDNMT1) in complex with hemi-methylated DNA (PDB: 4DA4) based on its reference sequence (NP_001124295.1).[10] mDNMT1 shares an 85% sequence similarity with the hDNMT1 reference sequence (**Figure A2.1 in Appendix A2**). The DNA sequence was then modified

to match the sequence used experimentally in determining DNA methylation rates. Each oxidized form of 5-methylcytosine (hmC, fC, caC) was subsequently modeled within the DNA template.

Molecular Dynamics

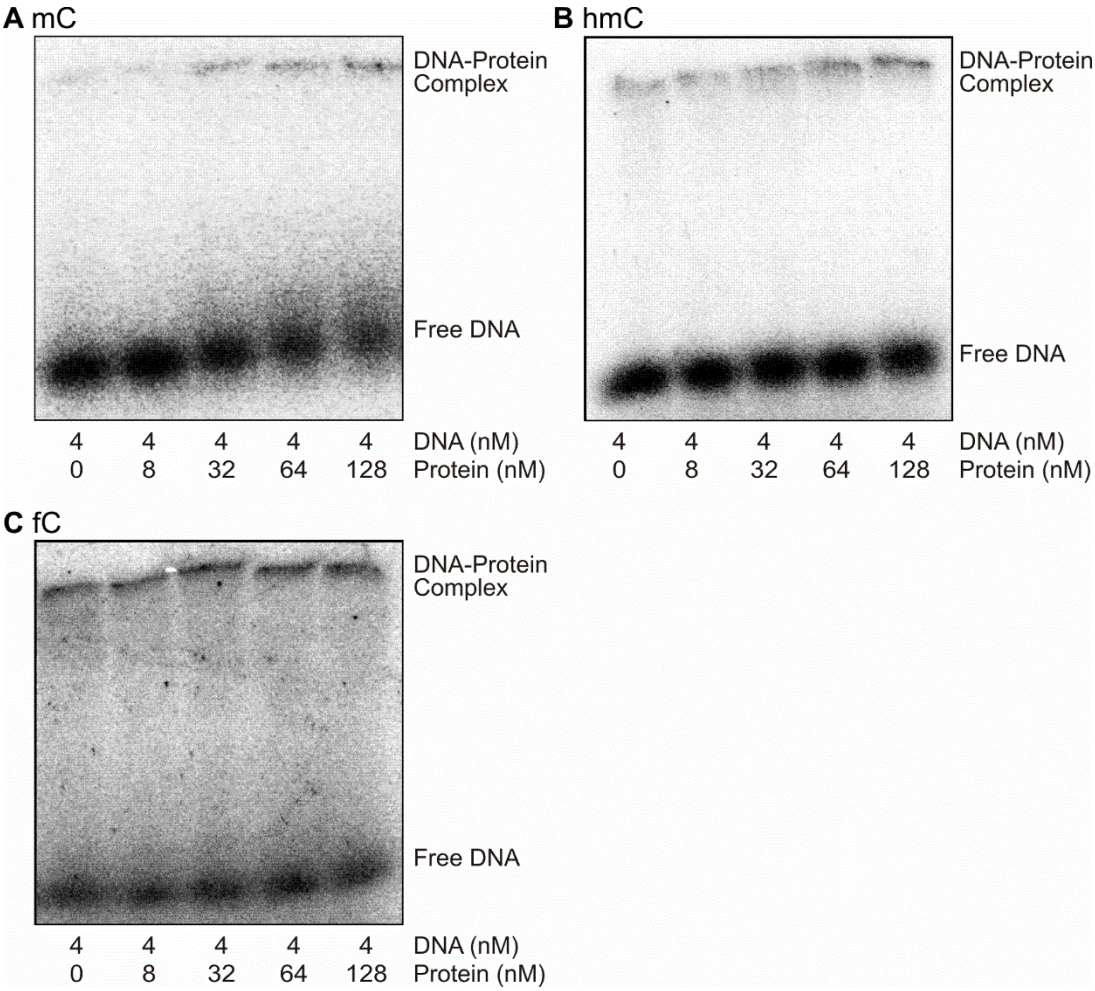
Desmond[269] was used to simulate each oxidized forms of 5-methylcytosine (hmC, fC, caC) within the modeled DNA template. As reported previously, each of the modeled hDNMT1 – DNA complexes was subjected to standard protein preparation protocols.[20] DNMT1-DNA complexes containing MeC or caC at a central CpG site were solvated with a 15 Å buffer region from its outer edge inside a rectangular box of TIP3P explicit solvent model.[270] 150 mM Na^+ and Cl^- counter ions were added to electroneutralize the final system. Each MD simulation was carried out using Desmond with default protocol for initialization, followed by 100 ns of unrestrained production simulation run under isothermal isobaric (NPT) conditions at 310 K and 1 atm with the OPLS3 force field. The long-range electrostatic interactions were evaluated by the Particle-Mesh Ewald method under periodic boundary conditions with a dielectric constant of 1. The stability of the protein-DNA complex was assessed by evaluating the protein C_αRMSD and DNA backbone phosphorus RMSD (P_RMSD)[271] with respect to the minimized starting structure.

3.3 Results

3.3.1 Kinetics of DNMT1 mediated cytosine methylation in the presence of MeC and its oxidized forms

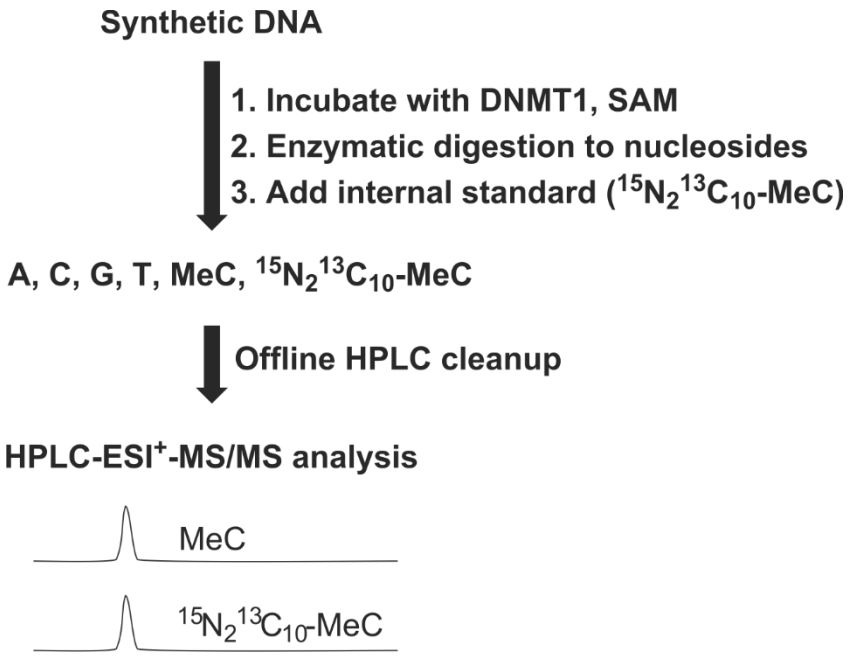
To examine the kinetics of DNMT1 mediated cytosine methylation in the presence of MeC and its oxidized variants, synthetic DNA duplexes (5′-AGCTTATCGCAGC XG GCGCGAATCTGA-3′) containing a single MeC, hmC, fC, or caC residue (X) at the central CpG site were prepared (Table 1). In the resulting double stranded DNA substrates (**Table 3.1**). In the resulting double stranded DNA substrates (**Table 3.2**), a single centrally located CpG site contains C (negative control), MeC (positive control), hmC, fC, or caC opposite unsubstituted cytosine. The ability of these 27-mer duplexes to serve as DNMT1 substrates was established by electrophoretic gel mobility shift (EMSA) assays, in which radiolabeled DNA duplexes were incubated with increasing amounts of human Δ580-DNMT1 protein (0-128 nM), followed by separation on 4% non-denaturing polyacrylamide gel to detect DNA-protein complexes[266] (**Figure 3.2**). An electrophoretic mobility shift characteristic of the formation of DNA-protein complexes was observed, consistent with DNMT1 binding to synthetic DNA duplexes.

Figure 3.2: Gel shift assays to detect Δ580-DNMT1 interaction with DNA duplexes containing MeC (A), hmC (B), and fC (C). No shift was observed for DNA containing C or caC. DNA sequence is provided in **Table 3.2**.



To establish the kinetics of DNMT1-mediated methyl transfer in the presence of MeC and its oxidized forms, a mass spectrometry based assay developed in our laboratory was employed (**Scheme 3.2**).[20] Following *in vitro* methylation reactions in the presence of human recombinant DNMT1 and S-adenosylmethionine cofactor, the amounts of newly formed MeC were determined by capillary HPLC-ESI$^+$-MS/MS using $^{13}C_{10}{}^{15}N_2$-mC internal standard.[20] In brief, DNA was spiked with known amounts of $^{13}C_{10}{}^{15}N_2$-mC and enzymatically digested to 2′-deoxynucleosides, which were analyzed by HPLC-ESI$^+$-MS/MS as reported previously.[20] Steady-state kinetic parameters (K_m and V_{max}) for methyl transfer reaction were determined by plotting the methylation velocity at a particular substrate concentration and using non-linear regression to fit the data to the Michaelis Menten equation.[20] The methylation velocity was calculated using Equation 1 as shown in the supplementary **Scheme 3.3**.

Scheme 3.2: Experimental scheme for the HPLC-ESI$^+$-MS/MS analysis of MeC.



Scheme 3.3: Example calculation for methylation velocity from LC-MS/MS peak areas using equation 1. Where A_{AN} is the measured area of the analyte, A_{IS} is the measured area of the internal standard, C_{IS} is the concentration of internal standard spiked, v is the final reaction volume in microliters, and t is the time of the reaction in minutes.

$$Velocity = \frac{(A_{AN}/A_{IS}) * C_{IS}}{[v * t * 1x10^{-6}]} = \frac{(A_{AN}/A_{IS}) * 1.33\ pmol}{25\ uL * 15\ \text{min} * 1x10^{-6}\frac{L}{uL}} = (A_{AN}/A_{IS})[3.55\ \frac{nM}{min}]$$

We found that DNMT1 methylation kinetics at CpG sites was strongly affected by the oxidation status of MeC in the opposite strand. The highest value of V_{max} (190 x 10^{-11} M/min) was observed for DNA duplexes containing MeCG dinucleotide (**Figure 3.3A**), which is similar to previous reports (27.2 – 163 x 10^{-11} M/min) that also observed a large effect of sequence context on methyl transfer rates.[272] Much lower rates of methyl transfer were observed for CpG sites containing hmC (41 x 10^{-11} M/min), fC (11.0 x 10^{-11} M/min), and caC (0.77 x 10^{-11} M/min) (**Table 3.3**). In general, DNMT1-mediated methylation rates decreased upon oxidation of MeC in the opposite strand (**Figure 3.3A, Table 3.3**). Indeed, the V_{max} values for methylation of cytosines placed opposite hmC, fC, and caC were 4-, 17-, and 240-fold lower than for MeC containing DNA, respectively (**Table 3.3**). The K_m values for MeC, hmC, fC, and caC were 28, 20, 13, and 1.1 nM respectively (**Table 3.3**). Catalytic efficiencies (V_{max}/K_m) for methyl transfer to cytosine residues opposite C, MeC, hmC, fC, and caC in CpG dinucleotides were calculated as 0.56, 6.7, 2.1, 0.85, and 0.71 ($x10^{-2}$ $min^{-1}M^{-1}$) respectively (**Table 3.3**). Overall, our results indicate that the ability of MeC to direct DNMT1-mediated maintenance methylation of CpG sites is reduced upon its oxidation to hmC, fC, and caC. This reduced maintenance methylation activity is likely to lead to passive DNA demethylation.[273] In contrast, DNMT1 efficiency was only weakly affected by local sequence context, with V_{max}/K_m values of 8.1, 12, 6.2, and 8.4 ($x10^{-2}$ $min^{-1}M^{-1}$) for MeCGG, MeCGC, MeCGA, and MeCGT, respectively (**Figure 3.3B, Table 3.4**).

Figure 3.3: Michaelis-Menton plots for full-length DNMT1 mediated methylation for DNA duplexes containing central CG, MeCG, hmCG, fCG, and caCG dinucleotides DNA duplexes were incubated with hDNMT1 and SAM for 15 min at 37 °C. After quenching, DNA was digested to nucleosides, spiked with $^{13}C_{10}{}^{15}N_2$-mC and analyzed for MeC by LC-MS/MS. The methylation velocity was plotted against DNA concentrations. (**A**) Influence of oxidation state of the methyl group on the rates of maintenance methylation in DNA duplexes 5′-AGCTTATCGCAGC XG GCGCGAATCTGA-3′ (X =C, MeC, hmC, fC, or caC). (**B**) Influence of 3'-neighboring nucleobase on methyl transfer rates for DNA duplexes of the sequence (5′-AGCTTATCGCAGC ^{Me}CGX CGCGAATCTGA-3′ where X = C, G, T, or A).

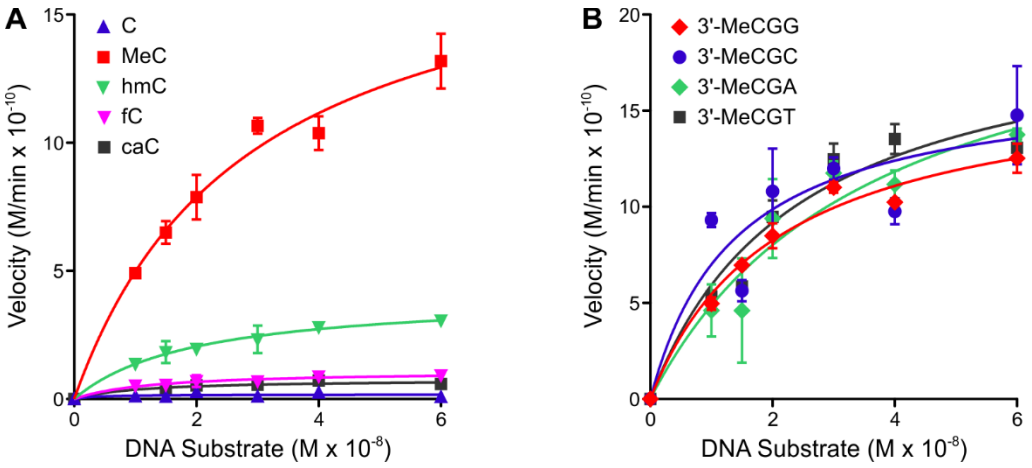


Table 3.3: Kinetic parameters of the DNMT1 mediated methylation of 27-mer duplexes containing modified cytosines.[a]

DNA Duplex	V_{max} ($x10^{-11}min^{-1}$)	K_m (nM)	V_{max}/K_m ($x10^{-2}\ min^{-1}M^{-1}$)
(-)C	1.8 ± 0.61	3.2 ± 8.6	0.56 ± 1.5
(-)MeC	190 ± 18	28 ± 5.7	6.7 ± 1.5
(-)hmC	41 ± 5.1	20 ± 6.2	2.1 ± 0.69
(-)fC	11 ± 1.1	13 ± 4.0	0.85 ± 0.33
(-)caC	0.77 ± 0.19	1.1 ± 1.0	0.71 ± 0.67

[a] The V_{max} and K_m values were determined via nonlinear regression using data from three or more individual points. The ranges in V_{max} and K_m are the standard error for regression analysis. Error was propagated for V_{max}/K_m using the equation: $\frac{Dc}{c} = \sqrt{\left(\frac{Da}{a}\right)^2 + \left(\frac{Db}{b}\right)^2}$; where a, b, and c are V_{max}, K_m, and V_{max}/K_m respectively.

Table 3.4: Kinetic parameters for DNMT1 mediated methylation of DNA sequences with varied sequence context.[a]

DNA Duplex	V_{max} ($x10^{-11} min^{-1}$)	K_m (nM)	V_{max}/K_m ($x10^{-2}\ min^{-1}M^{-1}$)
MeCGT	200 ± 20	24 ± 5.4	8.4 ± 2.0
MeCGA	230 ± 40	36 ± 12	6.2 ± 2.4
MeCGC	170 ± 24	14 ± 5.8	12 ± 5.3
MeCGG	170 ± 10	21 ± 2.9	8.1 ± 1.2

[a] The V_{max} and K_m values were determined via nonlinear regression using data from three or more individual points. The ranges in V_{max} and K_m are the standard error for regression analysis. Error was propagated for V_{max}/K_m using the equation: $\frac{Dc}{c} = \sqrt{\left(\frac{Da}{a}\right)^2 + \left(\frac{Db}{b}\right)^2}$; where a, b, and c are V_{max}, K_m, and V_{max}/K_m respectively.

3.3.2 Molecular modeling of DNMT1-DNA complexes containing oxidized forms of MeC

To establish the structural origins of reduced DNMT1 activity in the presence of oxidized forms of MeC, molecular models of DNMT1-DNA complexes were considered. For this purpose, homology models of the productive hDNMT1 complex with DNA duplexes containing MeC, hmC, fC, and caC were created (**Figure 3.4**). Molecular modeling of hDNMT1 protein was carried out using the published crystal structure of mDNMT1 in complex with hemi-methylated DNA.[10] The associated DNA duplex was modeled to reflect the sequence employed in our experimental studies (**Table 3.2**). The homology model was based upon the sequence alignment of hDNMT1 with the published crystal structure of mDNMT1 (PDB: 4DA4, **Appendix A2: Figure A2.1**). Alignment of the hDNMT1 reference sequence with the sequence of mDNMT1 reveals an 85% identity. Importantly, the residues making up the hydrophobic binding pocket of the TRD (C1501, L1502, W1512, and M1535) are conserved between mDNMT1 and hDNMT1 (**Figure 3.5**).

Figure 3.4. Homology model of the productive human DNMT1 – DNA complex. Homology modeling was performed using the Schrödinger modeling suite package and the crystal structure of mouse DNMT1 in complex with hemi-methylated DNA (PDB: 4DA4)[10] coupled with the reference sequence of hDNMT1.

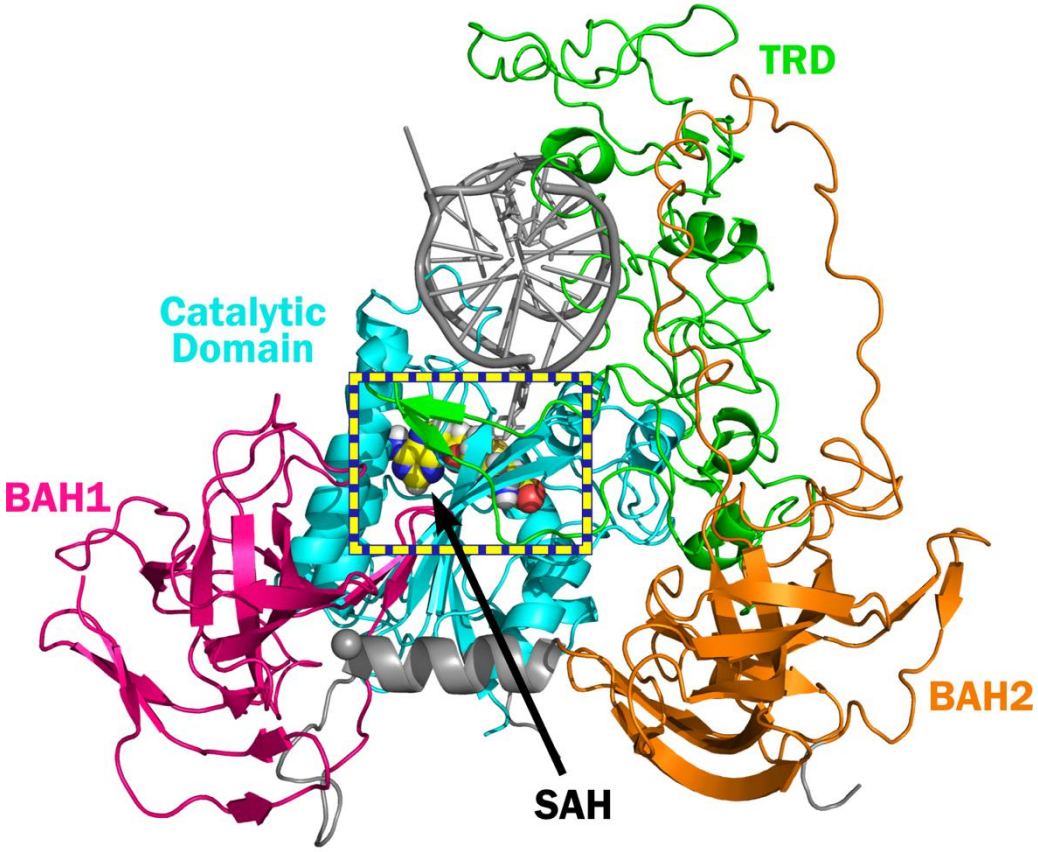
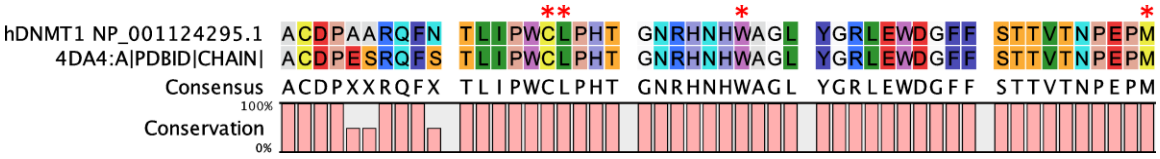


Figure 3.5. Sequence alignment of the target recognition domain. Alignment of the published crystal structure of mDNMT1 (PDB: 4DA4)[10] with the hDNMT1 reference sequence (NP_001124295.1) demonstrates that the residues making up the hydrophobic binding pocket of the TRD (C1501, L1502, W1512, and M1535) are conserved in both mDNMT1 and hDNMT1. These conserved residues are starred below.



In order to examine the structural effects of TET-mediated oxidation of MeC on DNA-protein binding, MeC was sequentially replaced with hmC, fC, and caC. Molecular dynamics (MD) simulations of hDNMT1 in complex with DNA containing either MeC, hmC, fC, or caC were performed to determine how oxidized forms of MeC influence the recognition of hemi-methylated CpG sites in DNA by the TRD of DNMT1. The TRD of human DNMT1 contains a hydrophobic binding pocket consisting of M1535, C1501, L1502 and W1512.[10] This pocket harbors the 5-methyl group of MeC and is involved in the recognition of the hemi-methylated MeCpG in productive hDNMT1-DNA complexes.[10]

MD simulations were carried out using the previously developed homology model of hDNMT1 in complex with hemi-methylated DNA.[20] The root-mean-square deviation (RMSD) was used to determine the stability of the protein and DNA structures over the MD simulation. The RMSD is a similarity measure widely used in the analysis of macromolecular structures and dynamics as it measures the total structural deviation from the starting position. The stability of the enzyme-DNA complex for each modification is demonstrated by the RMSD of the protein backbone (Cα RMSD) and the RMSD of the DNA phosphorous backbone (**Figure 3.6A and 3.6B**). The rise of the RMSD represents the equilibration from the coordinates of the initial model then the RMSD of MeC and its oxidized forms remains stable over the 100 ns of simulation. Monitoring the interactions of MeC and its oxidized forms with amino acid residues within the hydrophobic pocket of the TRD, we observed an iterative increase in distance between the oxidized forms of MeC

and DNMT1 residues including Cys1501, Leu1502, and Met1535 (**Table 3.5**, **Figure 3.7**, **Figure 3.8**). This suggests that oxidation to caC may lead to a conformational change to the unproductive mode of binding. Our results demonstrate that as MeC is oxidized to hmC, fC, and caC, the increased size and hydrophilicity of the C-5 substituent induces a spatial displacement of the oxidized MeC from the TRD binding pocket, disrupting the hydrophobic interactions between DNMT1 and DNA and leading to a loss of enzymatic activity. Our model supports the experimentally observed trend for hDNMT1 enzymatic activity that caC leads to the largest perturbation of the TRD, causing a significant loss in activity. (**Figure 3.3A, Table 3.2**).

Table 3.5: Average distance (Å) between MeC, hmC, fC or caC and residues in the TRD as calculated by the molecular dynamics simulation. Distances are averaged from 50-100 ns.

Residue	MeC distance (Å)	hmC distance (Å)	fC distance (Å)	caC distance (Å)
Cys1501	4.44 ± 0.49	6.47 ± 0.47	5.42 ± 0.46	10.12 ± 0.39
Leu1502	4.01 ± 0.36	5.47 ± 0.51	4.39 ± 0.35	8.85 ± 0.47
Met1535	4.39 ± 0.34	4.39 ± 0.30	5.88 ± 0.49	9.34 ± 0.75

Figure 3.6. Stability of the DNA-Protein complex. **A.** Cα RMSD demonstrates stability of the protein backbone during the MD simulation. **B.** The RMSD of the DNA phosphorus backbone shows the stability of the DNA duplex with respect to the starting minimized structure.

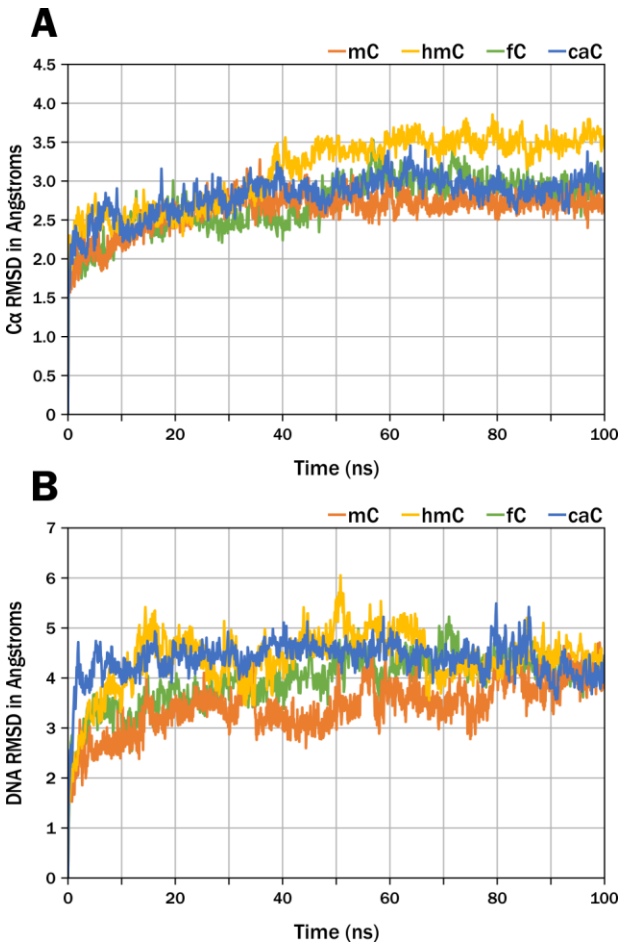


Figure 3.7. Molecular dynamics (MD) simulations demonstrate an incremental spatial displacement of oxo-mC from the TRD hydrophobic binding pocket. **A**. Residues Cys1501, Leu1502, and Met1535 make up the target recognition domain (TRD) and harbor the methyl group of MeC, providing the specificity of DNMT1 for hemi-methylated DNA. The MD simulations quantify the displacement of the oxidized forms of MeC from these residues in the TRD: **B**. Cys1501 **C**. Leu1502 **D**. Met1535

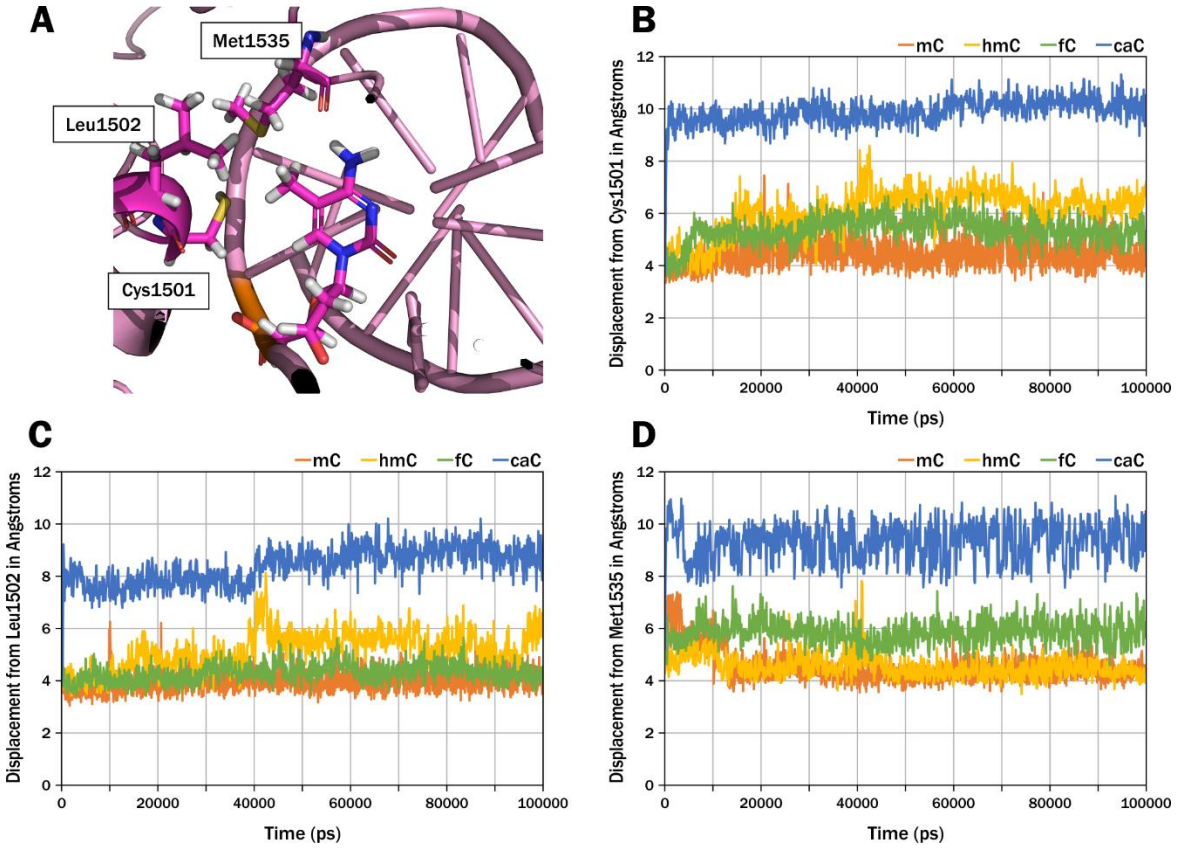
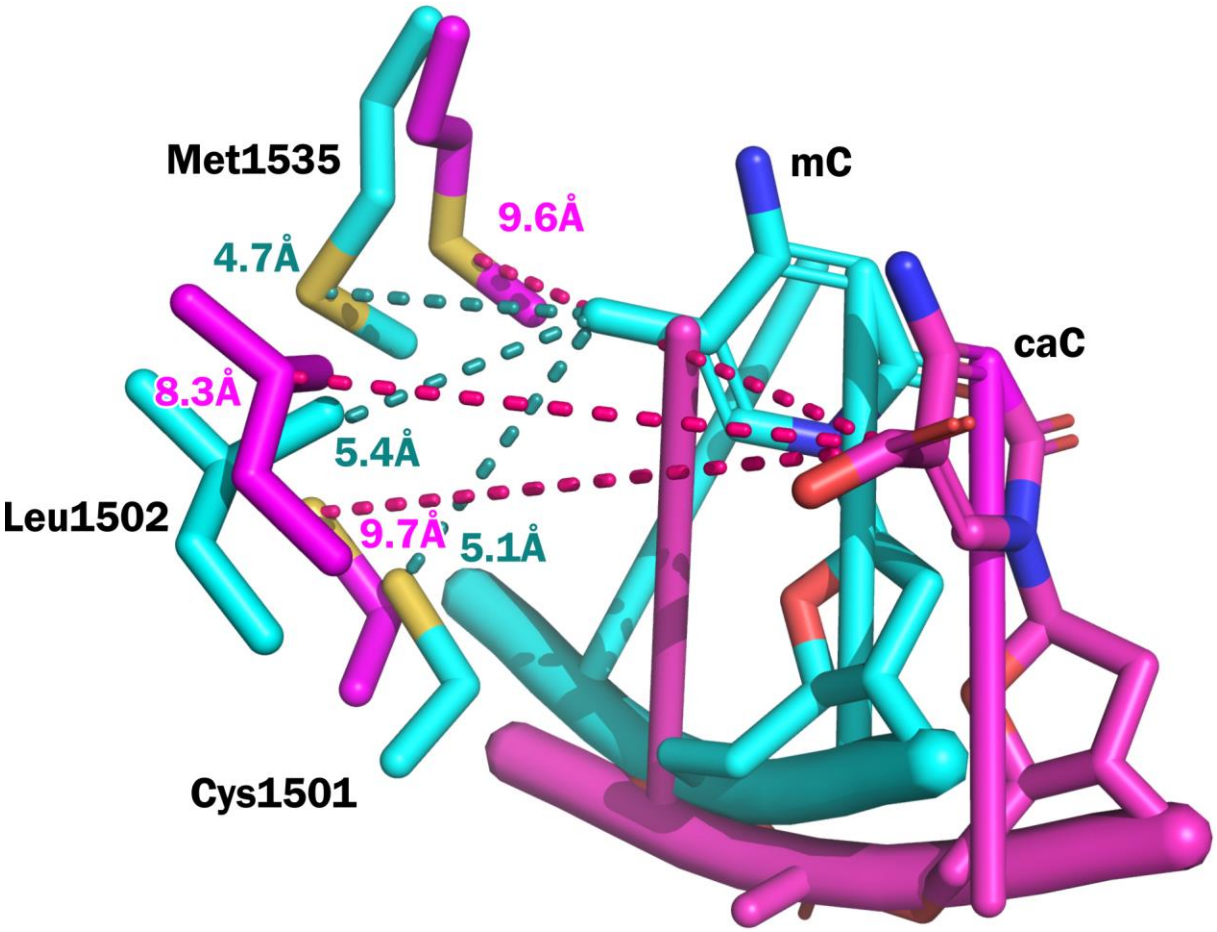


Figure 3.8. Overlay of MeC and caC structures after 100 ns MD simulation. The MeC structure is shown in cyan and the caC structure is shown in magenta. Distances between key residues which make up the TRD (Cys1501, Leu1502, Met1535) and the C5 position of each base are shown in dashed lines (mC: cyan, caC: magenta). The increased distances between caC and the residues in the TRD suggest a structural change in the DNMT1-DNA mode of binding.



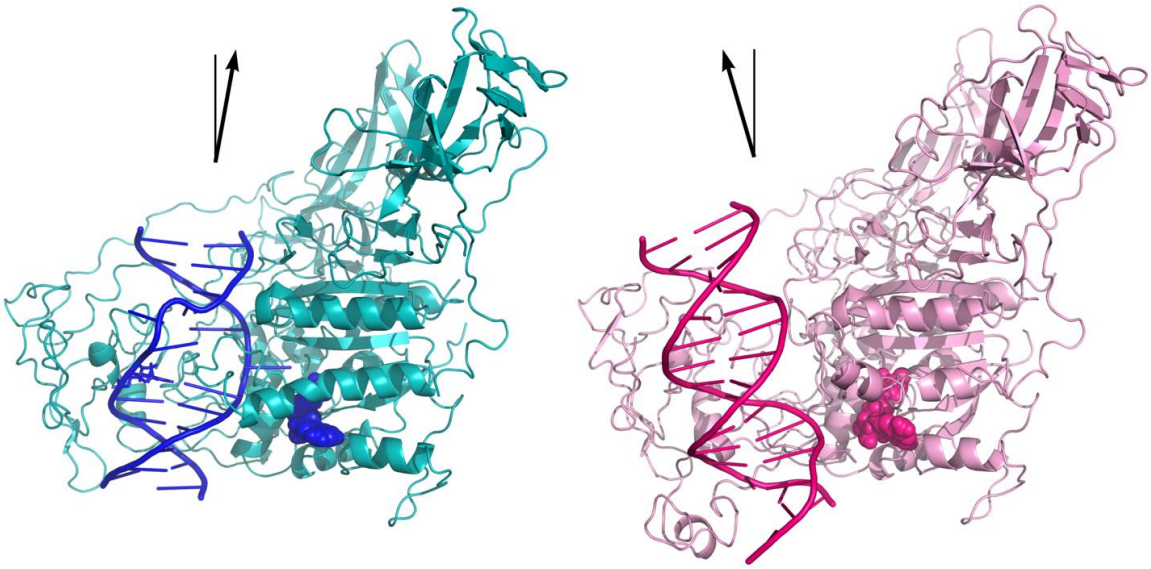
3.4 Discussion

Epigenetic DNA methylation marks (mC) must be removed as part of normal development,[274] neuronal plasticity,[275] and memory formation, resulting in chromatin remodeling and gene reactivation.[275] DNA demethylation can be accomplished via passive or active DNA demethylation processes.[276] Active demethylation is mediated by the iterative oxidation of MeC to hmC, fC, and caC by Ten Eleven Translocation (TET) proteins, followed by excision of fC and caC by TDG and their replacement with C via base excision repair mechanism.[13, 15-16, 263] Passive demethylation occurs when DNMT1 fails to methylate hemi-methylated DNA sequences generated during DNA replication.[5] Recent studies have shown that oxidized forms of MeC may participate in passive DNA demethylation by reducing the activity of maintenance methyltransferase (DNMT1).[19, 233] However, the structural origins and the mechanistic details for reduced DNMT1 activity in the presence of oxidized MeC variant remained unknown.

Two distinct modes of DNMT1-DNA binding are known. The protein adopts an unproductive mode of binding in complex with unmethylated DNA (PDB: 3PTA) and a productive mode of binding when bound to hemi-methylated (PDB:4DA4) DNA (**Figure 3.9**).[10, 20, 244] In the unproductive DNMT1-DNA complex, the double stranded DNA retains its base pairing, and the auto inhibitory mechanism described earlier prevents DNMT1 from performing *de novo* DNA methylation and directs its activity to hemimethylated sites.[244] In the productive DNMT1-DNA complex formed with CpG sites containing a single MeC, both the enzyme and its DNA substrate undergo a large conformational

change, which is initiated by MeC binding in a hydrophobic segment within the TRD (**Figure 3.9**).[10] This recognition of MeC by the TRD results in the insertion of amino acid sidechains from the catalytic and recognition domains of DNMT1 into both grooves of the DNA.[10] This productive mode of binding undergoes a local melting of the DNA duplex, rotating the target cytosine out of the DNA helix into the catalytic pocket and allowing for methyl transfer to take place.[10] The side chain of Met1235 inserts into the DNA from the minor groove and occupies the space vacated by the target cytosine.[10]

Figure 3.9. Crystal structures of DNMT1 bound to methylated DNA (teal, PDB: 4DA4)[10] or unmethylated DNA (pink, PDB: 3PTA)[244] reveal a productive and unproductive mode of binding, respectively. A shift in angle of binding highlighted by a black arrow. Cofactor, SAH, is shown as spheres.



The enzymatic mechanism of DNMT1 enzyme is well understood and involves nucleophilic addition of a cysteine in the active site to the C6 position of cytosine, followed by methyl transfer from the SAM cofactor to the C5 position.[10] Following methyl transfer, an excess proton from the C5 position is abstracted, and the covalent bond between the enzyme and cytosine base is cleaved to liberate the methylated DNA (**Scheme 2.7** in Chapter II of this Thesis).[261-262] Nucleophilic addition of the cysteine residue of DNMT1 to DNA is fast and reversible, while the following methyl transfer is the rate-limiting step during the formation of MeC.[277]

The purpose of the present work was to establish kinetic parameters for DNMT1 mediated methylation in the presence of MeC, hmC, fC, and caC and to elucidate the structural mechanisms of their effects on methyl transfer kinetics. A novel mass spectrometry based assay developed in our laboratory was used to follow the kinetics of methyl transfer.[20] We found that the rates of methyl transfer drastically decreased as the oxidation state of the methyl group on MeC increased from MeC to hmC, fC, and caC (V_{max}, 190, 41, 11, and 0.77 $x10^{-11}min^{-1}$, for MeC, hmC, fC, and caC respectively – see **Figure 3.3, Table 3.3**). Taken together with previous reports by Ji *et al*.[233] and Valinluck *et al*.,[19] our results show that unlike MeC, its oxidized forms fail to effectively direct DNMT1 enzyme to methylate the cytosine in the opposite strand. The V_{max}/K_m values also decreased according to the oxidation status hmC > fC > caC, indicating that the overall efficiency of enzymatic methylation was reduced. The catalytic efficiency for methyl transfer in the presence of for hmC, fC, and caC decreased 3.3-, 7.9-, and 9.5-fold relative

to MeC (**Table 3.3**). In addition, MeC oxidation to hmC, fC, and caC leads to lower K_m values for DNMT1-DNA binding (**Table 3.3**), indicative of the formation of tightly bound unproductive complexes. Indeed, previous reports by Pradhan *et al*. demonstrated that DNMT1 binds unmethylated DNA with higher affinity than hemi-methylated DNA.[266]

In contrast, local nucleotide sequence context had a minimal effect on methylation transfer kinetics (**Figure 3.3B, Table 3.4**) When the 3′ neighboring base was altered (MeCGX), this did not change the kinetic parameters for methyl transfer. The efficiency of methyl transfer was slightly higher in MeCGC context (**Figure 3.3B, Table 3.4**). This of interest because MeC is commonly found in promoter CpG islands of inactive genes.[278-279]

In our earlier study, extending the aliphatic side chain on C-5 of cytosine beyond methyl (5-ethyl-dC, 5-propyl-dC) resulted in a loss of DNMT1 maintenance methylation.[20] In that study, V_{max} and K_m values for MeC-containing DNA were determined as 9.6 x 10^{-2} nM/min and 21.8 nM using the sequence 5′-CGCGGA[mC]GCGGGTGCCGGG-3′.[20] As the length of the C5-alkyl chain increased, DNMT1-DNA binding via the TRD was disrupted, causing the enzyme to adopt an unproductive mode of binding to DNA.[20] Specifically, when the C5 alkyl chain on cytosine was extended from methyl to ethyl, we observed a 4-fold loss in V_{max}, and a similar efficiency of DNMT1-mediated methyl transfer (**Table 2.3 in Chapter II of this thesis**).[20] Further increase of the C-5 substituent size to propyl completely abolished methylation.[20] Since 5-ethyl-dC and hmC are of comparable size, but DNMT1 activity is 2-3 fold less efficient for hmC-containing

duplexes, this suggests that in addition to steric effects, oxidation of the methyl group of MeC leads to a loss of hydrophobic interactions with the TRD of the protein (**Table 3.3 and Table 2.3 in Chapter II of this thesis**).[20]

To examine the structural basis for reduced DNMT1 methylation activity in the presence of oxo-mC, a computational model of hDNMT1-DNA complex was developed. We found that as MeC was oxidized to hmC, fC, and caC, hydrophobic interactions responsible for DNMT1 recognition of hemimethylated CpG sites in DNA were disrupted. The increased distances between caC and the residues in the TRD (**Figure 3.7**) suggest a structural change in the DNMT1-DNA mode of binding. In the presence of oxo-mC, the loss of key hydrophobic interactions prevents the formation of a productive DNA-protein complex and instead DNMT1 forms tightly bound unproductive DNMT1-DNA complexes characterized by lower K_m value (**Figure 3.8, Figure 3.9, Table 3.3**).[10, 20, 266] Overall, our results confirm that oxidized forms of MeC participate in passive DNA demethylation and provide further kinetic and structural details for this important epigenetic process.

IV. EPIGENETIC EFFECTS OF INFLAMMATION AND EXPOSURE TO TOBACCO CARCINOGEN NNK IN THE A/J MOUSE MODEL OF SMOKING-INDUCED LUNG CANCER

Adapted from:

Seiler, C. L.; Song, J.M.; Kotandeniya, D.; Chen, J.; Kono, T.; Sarver, A.; Colwell, M.; Upadhyaya, P.; Ren, Y.; Faulk, C.; De Flora, S.; Chen, Y.; Kassie, F.; and Tretyakova, N. Y.; Epigenetic Effects of Inflammation and Exposure to Tobacco Carcinogen NNK in the A/J Mouse Model of Smoking-Induced Lung Cancer,

Manuscript in preparation

This work was performed in collaboration with Dr. Jung Min Song, Dr. Delshanee Kotandeniya, Jianji Chen, Dr. Thomas Kono, Dr. Aaron Sarver, Mathia Colwell, Dr. Pramod Upadhyaya, Yanan Ren, Dr. Chris Faulk, Dr. Silvio De Flora (University of Genoa Italy), Dr. Yue Chen, Dr. Fekadu Kassie, under the direction of Dr. Natalia Tretyakova. Dr. Delshanee Kotendeniya performed initial method development and quantitation of MeC, hmC, and fC in 2 week treated mice. Dr. Jung Min Song carried out the qRT-PCR for 9, 22, and 44-week mouse treatments, methylation specific PCR, and animal treatments under the direction of Dr. Fekadu Kassie. Jianji Chen and Dr. Yue Chen measure histone acetylation levels. Christopher Seiler and Mathia Colwell performed the bisulfite pyrosequencing for DNA from the 6-week study. Dr. Pramod Upadhyaya provided NNK for animal treatment. Yanan Ren performed statistical calculations for the LC-MS/MS data.

Dr. Chris Faulk designed the pyrosequencing assays for mTet1 and mAhrr. Dr. Silvio De Flora provided DNA from mice exposed to cigarette smoke. Christopher Seiler finalized and validated the HPLC-MS/MS methodology for epigenetic DNA marks and applied it to 6 weeks treated, and tumor samples. Christopher Seiler performed oxoBS-Seq and conducted data analysis with the guidance of Dr. Aaron Sarver. Dr. Tom Kono and Dr. Aaron Sarver performed bioinformatics support for oxBS-Seq data. Internal standards for fC and caC were synthesized by Christopher Seiler. Christopher Seiler and Natalia Tretyakova wrote the manuscript.

4.1 Introduction

Lung cancer is responsible for 30% of all cancer deaths worldwide and is expected to kill 154,050 Americans this year, with over 80% of cases directly attributable to smoking.[101] Cigarette smoke contains over 60 known carcinogens including tobacco specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), which forms promutagenic DNA adducts and induces cancer driving mutations in the *K-ras* protooncogene.[104] In addition to genetic alterations, epigenetic changes such as deregulated DNA methylation, altered histone acetylation, and aberrant micro RNA expression are common in smoking-induced lung tumors.[3]

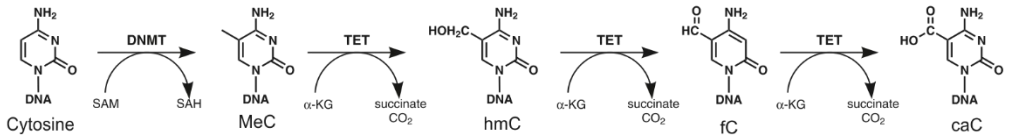
Aberrant DNA methylation patterns in malignant cells result in silencing of tumor suppressor genes, activation of protooncogenes, and decreased chromosomal stability.[280] Smoking-induced lung tumors exhibit elevated methylation of tumor suppressor genes such as MGMT, hMLH1, p16, RASSF1A, DAPK1, and RAR-β.[107-108] These "epimutations" cooperate with genetic alterations to drive the malignant lung tumor phenotype.[281] However, the mechanistic origins and the timing of epigenetic changes associated with smoking have remained largely unknown, limiting our understanding of lung cancer etiology and hindering the development of novel treatments.

Chronic inflammation plays a central role in the pathogenesis of lung cancer.[282] Smoking induces neutrophilic inflammation and reduces mucociliary clearance via several mechanisms, including exposure to endotoxins present in tobacco smoke (e.g. lipopolysaccharide (LPS), 2,400 ng/pack).[105] When administered intranasally to laboratory

mice, LPS induces inflammatory response mimicking chronic obstructive pulmonary disease (COPD), a major risk factor for lung cancer development in smokers.[141] Sustained inflammation in the lung in humans can lead to COPD, increasing the risk of developing lung cancer in smokers.[283-284] Indeed, chronic exposure of A/J mice to LPS increases lung tumor size and multiplicity following treatment with NNK, supporting a role for inflammation in lung cancer etiology.[141] Due to the increased risk of lung cancer from chronic inflammation, the use of anti-inflammatory agents is a potential chemopreventative strategy.[140] Common non-steroidal anti-inflammatory agents (NSAIDs) have been shown to reduce the prevalence of colorectal cancers,[285] and may affect lung carcinogenesis.[286-288]

In the present work, we examined the dynamic changes in cytosine methylation, hydroxymethylation, formylation (**Figure 4.1**), histone acetylation, and gene expression patterns in lung tissues of A/J mice chronically exposed to cigarette smoke, NNK and LPS alone, or in combination. Our results reveal that many of the epigenetic changes characteristic for lung tumors can be observed soon after an inflammatory stimulus. Because epigenetic alterations are expected to be reversible, the use of epigenetic modulators may open a possibility for novel strategies for lung cancer chemoprevention and treatment.

Figure 4.1. Epigenetic DNA marks and enzymatic pathways involved in their formation. Formation of MeC from cytosine is catalyzed by DNMT enzymes. Oxidation of MeC to hmC, fC, and caC is carried out by TET dioxygenases and can lead to DNA demethylation.



4.2 Materials and Methods

Chemicals and Enzymes.

PDE I, PDE II, and DNase I were purchased from Worthington Biochemical Corp. (Lakewood, NJ), while calf intestinal alkaline phosphatase was from Sigma-Aldrich, (Madison, WI). Nanosep10K filters were acquired from Pall corp. (Port Washington, NY). O-(biotinylcarbazoylmethyl) hydroxylamine was obtained from Cayman Chemical (Ann Arbor, MI). Isotopically labeled D_3-hmC was purchased from Cambridge Isotope Labs (Cambridge, MA). $^{13}C_{10}$,$^{15}N_2$-5-methyl-2′-deoxycytidine was synthesized as previously described.[20] All other chemicals used were bought from Sigma-Aldrich (Milwaukee, WI) or Fisher Scientific (Fairlawn, NJ).

Animal Treatments

Female A/J mice were obtained from the Jackson Laboratory (Bar Harbor, ME) and housed in specific-pathogen-free animal quarters at Research Animal Resources, University of Minnesota Academic Health Center. All animal experiments were performed according to the U.S. National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals and was approved by the Institutional Animal Care and Use Committee, University of Minnesota.

Short Term Treatment of mice with NNK, LPS and NNK/LPS: Female A/J mice (6 weeks of age) were divided into 3 groups (3 animals per group). Mice in Group 1 were treated

intraperitoneally (IP) with NNK (25 mg/kg, in 0.3 ml physiological saline solution) every day for 3 – 9 days and 12.5 mg/kg NNK daily on days 10-15 days. For lipopolysaccharide (LPS) treatments, 8.3 µg of LPS was administered intranasally twice a week in the first week of the experiment, and 4.15 µg LPS was given once a week during the second week. Mice were euthanized in a CO_2 chamber the day after NNK treatment. Tissues (lung, brain, kidney) were harvested and stored frozen at -80 °C until analysis.

Chronic Treatment with NNK, LPS, and NNK/ LPS: Female A/J mice (6 weeks of age) were divided into four groups: control (N = 4), NNK (N = 5), LPS (N = 4), and NNK+LPS (N = 4). NNK was administered IP (50 mg/kg) twice a week during weeks 1 and 3, and once during week 5 for a total of 5 doses. LPS was given intranasally (4 µg/mouse) twice in week 1, and once a week during weeks 3, 4, and 5, for a total of 5 doses. Control mice were treated with physiological saline IP for the same duration of time. Mice were euthanized in a CO_2 chamber during the sixth week of treatment. Tissues (lung and brain) were harvested and stored at -80 °C until analysis.

Animal treatment for gene expression analyses: Six-week-old female A/J mice received a single dose of NNK (100 mg/kg, IP). Mice were intranasally treated with LPS once a week under isoflurane anesthesia for 9 or 22 weeks (5 µg LPS per dose in 50 µL of PBS) or for 44 weeks with reduced LPS (2 µg per dose in 50 µL of PBS). After 9, 22, or 44 weeks of treatment, the mice were euthanized with an over dose of carbon dioxide. Whole lung tissues were collected from mice treated for 9 weeks and tumors from 22 and 44-week treatments and kept at -80 °C.

Lung Tumor induction with NNK/LPS: Female A/J mice (7 weeks of age) were divided into three groups (20 per group). Group 1 was treated intraperitoneally (IP) with NNK (100 mg/kg, in 0.3 ml PBS) once a week for two weeks, group 2 was treated intranasally with 4 μg of LPS (in 50 μL of 1 X PBS) a week after the 2nd dose of NNK administration and subsequently treated with the same dose of LPS once a week until week 27. Mice in control group 3 were given physiological saline (0.3 mL, IP). Mice were euthanized in a CO_2 chamber. Tumors (≥ 100) were pooled from lung lobes of 4-5 mice in group 1 and 2 and stored frozen at -80 °C until analysis.

Cigarette smoke exposure treatment: This study was conducted at the University of Genoa, Italy. Newborn, female A/J mice were divided into two groups (4 per group). Group 1 was exposed to filtered air for the duration of the study, and group 2 was exposed to environmental cigarette smoke starting 12 hours after birth. Mice were euthanized in a CO_2 chamber. The lungs were isolated, and DNA was extracted from the left lung and stored at -20 ℃ until analysis.[140]

Synthesis of Stable Isotope Labeled Internal Standards

$^{13}C_{10}{}^{15}N_2$-5-formyl-2′-deoxycytidine: To the stirring solution of $^{13}C_{10}{}^{15}N_2$-5-methyl-2′-deoxycytidine (1 mg, synthesized previously in Chapter II of this thesis)[20] in a 1 M phosphate buffer at pH 7 was added sodium persulfate (16.4 mg). The mixture was placed in a preheated oil bath and stirred at 70 °C for 3 days. The reaction mixture was

purified by RP-HPLC using a Synergi Hydro-RP (250 mm x 4.6 mm x 4 µm, Phenomenex, Torrance CA) with a linear gradient of acetonitrile and water (3% to 5% over 25 min), with product eluting at 16.8 minutes.

$^{13}C_{10}{}^{15}N_2$-5-carboxyl-2′-deoxycytidine: $^{13}C_{10}{}^{15}N_2$-5-formyl-2′-deoxycytidine (57 µg, 210 nmol) was placed in an Eppendorf tube along with potassium monophosphate (16 mg, Sigma-Aldrich), isoamylene (82 µL, Sigma-Aldrich), sodium chlorite (16 mg, Sigma-Aldrich), water (180 µL), tetrahydrofuran (444 µL, Fisher Scientific), and t-butanol (800 µL, Fisher Scientific). The reaction was stirred for 2.5 hours at room temperature and then acidified with 400 µL of 1 N HCl. The reaction mixture was separated by RP-HPLC using a gradient of 0.1% formic acid in water and 0.1% formic acid in acetonitrile on a Synergi Hydro-RP column (250 mm x 4.6 mm x 4 µm, Phenomenex) with a linear increase from 3% to 5% acetonitrile and $^{13}C_{10}{}^{15}N_2$-cadC eluted at 7 min.

DNA Digestion and Enrichment of MeC, hmC, fC, and caC

Genomic DNA was extracted from lung, kidney, and brain tissues using IBI-Mini Genomic DNA Kit (IBI Scientific, Peosta IA) according to the manufacturer's protocol. DNA concentrations were determined by UV spectroscopy.

Genomic DNA (2-10 µg) was subjected to hydrolysis with PDE I (3.6 U, 5 µg), PDE II (3.2 U), DNase I (50U), and alkaline phosphatase (10 U) in 10 mM Tris HCl/15 mM $MgCl_2$ buffer (pH 7) at 37 °C overnight. The hydrolysates were spiked with $^{13}C_{10}{}^{15}N_2$-5-methyl-2′-deoxycytidine (1 pmol), 5-hydroxymethyl-d_2-2′-deoxycytidine-6-d_1 (900

fmol), $^{13}C_{10}{}^{15}N_2$-5-formyl-2′-deoxycytidine (500 fmol), and $^{13}C_{10}{}^{15}N_2$-5-carboxyl-2′-deoxycytidine (300 fmol) (internal standards for mass spectrometry synthesized in our laboratory)[20] and filtered through Nanosep 10K Omega filters (Pall Corporation, Port Washington, NY).

DNA hydrolysates were dissolved in 100 mM ammonium acetate buffer (pH 4.5) containing 100 mM aniline and 400 μM O-(biotinylcarbazoylmethyl) hydroxylamine (Cayman Chemical. Ann Arbor, MI) and allowed to react for 24 h to derivatize fC to biotinyl-fC.[14] The resulting mixture was dried, followed by offline HPLC to enrich for MeC, hmC, biotinyl-fC, and caC using an Atlantis T3 column (Waters, 4.6 x 150 mm, 3 μm) was eluted at a flow rate of 0.9 mL/min with a gradient of 5 mM ammonium formate buffer, pH 4.0 (A) and methanol (B). Solvent composition was changed linearly from 3 to 30% B over 15 min, increased to 80% over the next 3 min, maintained at 80% B for the next 2 min, and brought back to 3 % B. The column was equilibrated for 7 min. dC was quantified by HPLC-UV using calibration curves obtained by analyzing authentic dC standards. HPLC fractions corresponding to MeC, hmC, caC, and biotinyl-fC (7-8.6 min for both hmC and caC, 9-10.5 min for MeC, and 18.7-20.2 min for biotinyl-fC) were combined, dried, and analyzed by isotope dilution HPLC-ESI-MS/MS.

HPLC-ESI$^+$-MS/MS quantitation of MeC, hmC, and fC

Quantitation of MeC, hmC, and biotinyl-fC was performed using a Dionex Ultimate 3000UHPLC (Thermo Fisher, Waltham MA) interfaced with a Thermo TSQ

Vantage mass spectrometer (Thermo Fisher). Chromatographic separation was achieved on a Zorbax SB-C18 column (0.5 x 150 mm, 3 μm, Agilent) eluted at a flow rate of 15 μL/min with a gradient of 2 mM ammonium formate (A) and methanol (B). Solvent composition was maintained at 5% B for the first 3 min and linearly changed from 5 to 40% B for 7 min. Solvent composition was a returned to initial conditions (5% B), and the column was re-equilibrated for 4 min. Under these conditions, MeC and $^{13}C_{10}{}^{15}N_2$-MeC eluted at 4.1 min, both hmC and the internal standard d_3-hmC eluted at 3.5 min, while biotinyl-fC and its internal standard ($^{13}C_{10}{}^{15}N_2$-biotinyl-fC) eluted at 7.0 min. Quantitation was achieved by monitoring the transitions *m/z* 258.1 [M + H^+] → *m/z* 141.1 [M – deoxyribose + H^+] for hmC, *m/z* 261.2 [M + H^+] → *m/z* 145.1 [M – deoxyribose + H^+] for D_3-hmC, *m/z* 242.1 [M + H^+] → *m/z* 126.1 [M + H^+] for MeC, *m/z* 254.2 [M + H^+] → *m/z* 133.1 [M + H^+] for $^{13}C_{10}{}^{15}N_2$-MeC, *m/z* 569.1 [M + H^+] → *m/z* 453.3 [M – deoxyribose + H^+] for Biotinyl-5fC, *m/z* 581.2 [M + H^+] → *m/z* 460.4 [M – deoxyribose + H^+] for Biotinyl-$^{13}C_{10}{}^{15}N_2$-5fC. Mass spectrometer were determined by infusion of authentic standards. Typical MS settings were: a spray voltage of 2700 V, a sheath gas of 15 units, the declustering voltage was 5 V, and the ion transfer tube was maintained at 350 °C. The full-width at half-maximum (FWHM) was maintained at 0.7 for both Q1 and Q3. Fragmentation was induced using a collision gas of 1.0 mTorr and a collision energy of 10.3 V.

HPLC-ESI$^-$-MS/MS analyses of caC

caC was quantified in the negative ion mode by monitoring the transitions *m/z* 270.00 [M - H^+] → 109.97 [M – deoxyribose – CO_2 - H^+] for caC and *m/z* 282.00 [M - H^+] → 116.05 [M – deoxyribose – CO_2 - H^+] for $^{13}C_{10}{}^{15}N_2$-caC internal standard. Chromatographic separation was achieved on a Thermo Hypercarb column (0.5 x 100 mm, 3 μm, Thermo Fisher Scientific) eluted at a flow rate of 14 μL/min with a gradient of 2.5 mM ammonium bicarbonate (A) and acetonitrile (B). Solvent composition was increased from 10 to 15% B over the first 3 min and further to 60% B over 7 min. Solvent composition was returned to initial conditions over 1 min and re-equilibrated over 6 min. Under these conditions, caC and $^{13}C_{10}{}^{15}N_2$-caC eluted at 5.5 minutes. MS parameters were optimized to achieve maximum sensitivity. The mass spectrometer was operated in negative mode with an S-lens voltage of 72 V, a spray voltage of -3.0 kV, capillary temperature at 270 °C, declustering voltage of 7 V, and nitrogen as a sheath gas at 20 arbitrary units. CID was achieved at with collision energy of 18 V. Argon was used as a collision gas with a pressure of 1.0 mTorr. MS/MS analyses were performed with a scan width of 0.1 *m/z* and a scan time of 0.1.

HPLC-ESI$^+$-MS/MS method validation

Fixed amounts of isotopically labeled internal standards (1 pmol $^{13}C_{10}{}^{15}N_2$-MeC, 900 fmol d_3-hmC, and 500 fmol $^{13}C_{10}{}^{15}N_2$-fC) and increasing amounts of the corresponding unlabeled nucleosides (2 pmol – 200 pmol MeC, 100 fmol – 20 pmol hmC, and 1 fmol – 500 fmol fC) were spiked into 1 μg of commercial calf thymus DNA (Sigma). Samples

were processed by enzymatic hydrolysis, ultrafiltration, and offline HPLC as described above and subjected to HPLC-ESI-MS/MS analysis. The observed amounts of MeC, hmC, and fC were plotted against theoretical values followed by linear regression analysis (**Figure 4.2**).

The LOD and LOQ values of the new method were determined by spiking synthetic dsDNA (5 μg, sequence: 5′-AGCTTATCGCAGCCGGCGCGAATCTGA-3′) with increasing amounts of synthetic oligodeoxynucleotides each containing a single MeC, hmC, or fC residue (5′-AGCTTATCGCAGC**X**GGCGCGAATCTGA-3′, where X = MeC, hmC, or fC) and fixed amounts of internal standards (1 pmol $^{13}C_{10}{}^{15}N_2$-MeC, 900 fmol d_3-hmC, and 300 fmol $^{13}C_{10}{}^{15}N_2$-fC), followed by sample processing and capillary HPLC-ESI-MS/MS analysis as described as above. The LOD values were determined as the lowest analyte amounts that consistently produced signal-to-noise ratios above 3. The LOQ was defined as the minimum amount of analyte that produced a coefficient of variation less than 15% and a signal-to-noise ratio greater than 10.

To evaluate precision and inter-day and intra-day accuracy of the method, samples were processed as above and analyzed three times per day on three consecutive days. Accuracy was calculated for each analyte using the equation: A_m/A_a x 100%, where A_m is the measured amount of analyte and A_a is the amount of analyte added.

Histone acetylation analysis:

Lung tissues from control and treated mice were homogenized with a glass douncer on ice. Total histones were extracted from the lysates as previously described.[289] Histone proteins were acetylated with ($^{13}C_2$,d_3)-acetyl N-hydroxysuccinimide ester to block all unmodified lysines and then digested by trypsin.[290] The peptides were desalted with C18 Stage Tips (3M Corporation, St. Paul, MN) prior to analysis.

Tryptic peptides were analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry (nanoLC-ESI-MS/MS) using a Thermo Scientific Orbitrap Fusion mass spectrometer (Thermo Scientific, San Jose, CA) coupled to a Proxeon Easy nLC 1000 UPLC system (Thermo Fisher Scientific, Odense, Denmark). Each sample was re-suspended in HPLC buffer A (0.1% formic acid in water) and loaded onto an in-house packed C18 column (25 cm x 75 μm I.D.) packed with ReproSil-Pur Basic C18 beads (2.5 μm, Dr. Maisch GmbH). Peptides were eluted with a gradient of 5% to 15% B (0.1% formic acid in acetonitrile) over 26 minutes, then 15% to 35% HPLC buffer B over 16 minutes at 300 nL/min.

Tryptic peptides from histone proteins were analyzed using a FT survey scan from 300-1600 *m/z* at a resolution of FWHM 120,000 (at 200 *m/z*), followed by HCD MS/MS scans using the top speed mode (3 seconds per cycle) at a resolution of FWHM 15,000 (at 200 m/z) and the normalized collision energy at 35%. The targeted MS/MS data acquisition was achieved with an inclusion list for fully labeled histone tryptic peptides that covered known lysine acetylation sites. For each peptide, modification isomers with all possible combinations of light/heavy lysine acetylation (delta mass of 42.010565 and 47.036094

Da, respectively) at detectable charge states were considered in the inclusion list for targeted fragmentations.

MS data was searched against the Uniprot Mus musculus proteome database (http://www.uniprot.org) using MaxQuant search engine (v1.4.1.2) as previously described.[282-284] Heavy and light acetylation on lysine as well as methionine oxidation were included as variable modifications, with 6 ppm specified as the precursor mass error and 0.025 Da as the fragment mass error. All peptide spectra matches were filtered at 1% False Discovery Rate with a minimum Andromeda score cutoff of 40. The HPLC elution profile were manually evaluated to ensure accurate quantifications. Only peptides that were confidently identified were selected for stoichiometry analysis. Acetylation stoichiometries of specific sites were calculated using in-house developed scripts based on the extracted peak areas of each modification isomer and quantification of modification-specific fragment ions.[1] Statistical significance analysis of site-specific acetylation stoichiometry dynamics between control and treated samples was conducted using two-sided Student's t-test using SAS statistical software 9.3 (SAS Institute Inc., Cary, NC).

Methylation specific PCR (MSP) assay

Genomic DNA was isolated from mouse lung tissues or lung tumors using a QIAamp DNA mini kit (Qiagen, Valencia, CA). For bisulfite conversion, 0.5 μg of the isolated genomic DNA was treated with sodium bisulfite using an EpiTect Bisulfite kit (Qiagen, Valencia, CA).

Methylation specific nPCR of bisulfite treated DNA was performed using EpiTect MSP kit (Qiagen, Valencia, CA) using methylated and unmethylated primer sets listed in **Table 4.1**. PCR reaction conditions were performed as follows: 1 cycle at 95°C for 10 min; 40 cycles at 94°C for 15 s, 48°C for 30 s, and 72°C for 30 s; 1 cycle at 72°C for 10 min. The resulting PCR products were analyzed on 2% agarose gels after staining ethidium bromide. Three human NSCLC cell lines (A549, H1299 and H2009) and a mouse cancer cell line (MCS) derived from mouse lung tumors induced were used as positive control samples to indicate the correct fragment by MSP.

Table 4.1. MSP primer sequences employed in this study.

MSP primer	Forward primer (5′-3′)	Reverse primer (5′-3′)
mDAPK1		
Methylated	AGGAGTCGCGAGCGTAGC	CAACTATCGCGTACGC
Unmethylated	TGGGAGGAGTTGTGAGTGT	ACAACTATCACTTCATAC ACC
mRAR-BETA		
Methylated	TCGTGGTTTTTTTGTGCGGTTC	CAACATACAAAAAAAAAAACTGCGG
Unmethylated	TTGTGGATTTTTTTGTGTGGTTTG	CAACATACAAAAAAAAAAACTCACAA
mGATA2		
Methylated	ATTAGGTAGATAGGGCGTAGAGTTC	CTAACTATCTCTCGATTCCCGAC
Unmethylated	GATTAGGTAGATAGGGTGTAGAGTTTG	TTCTAACTATCTCTCAATTCCCAAC
mCDH13		
Methylated	TATTTGTTATGTAAAACGAGGGAGC	CAAATAAATCAACGACAACATCG
Unmethylated	TTTGTTATGTAAAATGAGGGAGTGT	CCAAATAAATCAACAACATCAC
mRUNX3		
Methylated	TGTAGTTATAAGATTTTTTAAGGGGTC	CACAAAATACAAAAAACCAACTCG
Unmethylated	GTAGTTATAAGATTTTTTAAGGGGTTGT	TCACAAAATACAAAAAACCAACTCA
hDAPK1		
Methylated	GGGATTTTAGTATATATTTCGGGAC	GAACTACCCTACCAAACCGA
Unmethylated	TTGGGATTTTAGTATATATTTTGGGAT	CAAACTACCCTACCAAACCA
hRAR-BETA		
Methylated	GGTTAGTAGTTCGGGTAGGTTTTATC	CCGAATCCTACCCCGACG
Unmethylated	TTAGTAGTTTGGGTAGGGTTTATT	CCAAATCCTACCCCAACA
hGATA2		
Methylated	CGGGTATTTTTTTGTTTTTTGC	TAACCTCGCTACCTTCCTAACG
Unmethylated	TTTTGGGTATTTTTTTGTTTTTTGT	CTAACCTCACTACCTTCCTAACACT
hCDH13		
Methylated	AAGAAGTAAATGGGATGTTATTTTC	AAAACCAATAACTTTACAAAACGAA
Unmethylated	TTAAAGTAAATGGGATGTTATTTTT	ACCAAAACCAATAACTTTACAAAACA
hRUNX3		
Methylated	GGTTTAGTTAATGAGTTAAGGTCGC	TCTAATAAATACGAAAACGACCGA
Unmethylated	TTTAGTTAATGAGTTAAGGTTGTGA	TCTAATAAATACAAAAACAACCAA A

Quantitative reverse transcription–PCR (qRT-PCR) analysis

Total RNA was extracted from frozen mouse lung or lung tumor tissues using the miRNeasy Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instruction. The purity and the integrity of total RNA were confirmed by Nanodrop UV-Spectrophotometer and the RNA was stored at -80 ℃ until later use. The first-strand complementary DNA was synthesized by using QuantiTect Reverse Transcription Kit (Qiagen, Valencia, CA) with one microgram of RNA in 20 μL reaction. The first-strand complementary DNA mixture was further diluted to 200 μL with RNase-free water and stored at -20 °C until use.

qRT–PCR was performed by Light Cycler 96 (Roche, Indianapolis, IN) using QuantiTect SYBR Green PCR Kit (Qiagen, Valencia, CA) and gene specific primers (**Table 4.2**). Twenty five nanograms of complementary DNA was added to a 20 μL reaction. The final concentration of each primer was 0.5 μM. For PCR amplification, a program of initial denaturation at 95 °C for 15 min, followed by 45 cycles consisting of denaturation at 94 °C for 15 s, annealing at 50 °C for 30 s and extension at 72 °C for 34 s was used. All samples were normalized to an internal control (β-actin). Comparative Ct method was used to assess the relative levels of gene expression. Values were expressed as relative units compared with vehicle control (mouse lung tissue) and the standard error.

Table 4.2. Primer sequences used for qRT-PCR.

Gene name	Forward primer (5′-3′)	Reverse primer (5′-3′)
mDAPK1	CCGCTGTCAACTACGACTTT	GTCCTGGATTGTCATCCTCTTC
mRAR-β	CCTCTGACTGACCTTGTGTTC	GGCGGTCTCCACAGATTAAA
mGATA2	GACGACAACCACCACCTTAT	TGCTGGACATCTTCCGATTC
mCDH13	CCTGACAAGCCATCTCCTAAC	GACATCCAATCCTGCCATATCT
mPRDM2	TAGGTCCCGTGTGTGTATCT	CTGCTTTCCCATCACTCTGT
mRUNX3	GAGTTTCACGCTCACAATCAC	GCCTTGGTCTGGTCTTCTATC
mRASSF1	GAGACACCCGATCTTTCTCAAG	CACTGAAACAGGACGCACTA
mβ-actin	ACTCTTCCAGCCTTCCTTCC	GTACTTGCGCTCAGGAGGAG
mTet1	AGATGGCTCCAGTTGCTTATC	CTTCCGTTGTGCATGTTGTG
mTet2	GTCCTGATGTGGCAGCTATT	TCCTCACTCGATCTCCGATATAC
mTet3	GAGTTCCCTACCTGCGATTG	TCCATGAGTTCCCGGATAGA

Methylation Analysis by Pyrosequencing

DNA isolated from mouse lung tissues (100 ng) was treated with bisulfite using EpiTect Bisulfite Kit (Qiagen, Frederick MD) according to the manufacturer's instructions. Bisulfite treated DNA was amplified by PCR with primers for the following genes: Ahrr, DAPK1, CDH13, Tet1, and Rassf1. The thermocycler protocol for amplification reactions consisted of 95°C for 30 sec, the optimal annealing temperature for 30 sec, and 72°C for 30 sec, each reaction completing 40 rounds of PCR. PCR product amplicon sizes were checked on a QIAxcel (Qiagen) prior to pyrosequencing to ensure the correct gene was amplified. Bisulfite converted DNA was prepared for pyrosequencing according to the instructions in the PyroMark assay kit (Qiagen, Frederick, MD), including three controls: "no template control" (NTC), bisulfite converted 100% and 0% methylated mouse DNA. Methylated mouse DNA controls were created by methylation of genomic mouse DNA with M.SssI. Pyrosequencing was carried out according the design files from Qiagen and the Qiagen Assay Design Software on the PyroMark Q96 (Qiagen). CpG methylation values were used for statistical analysis, and values defined as "failed" were discarded. Statistical significance was calculated in GraphPad Prism 6 using 2-way ANOVA.

4.3 Results

Development of quantitative HPLC-ESI-MS/MS methodology for MeC, hmC, fC, and caC

To allow for accurate quantitation of epigenetic DNA marks, we have developed an isotope dilution HPLC-ESI-MS/MS methodology for MeC, hmC, fC, and caC (**Figure 4.2**). Our approach involves enzymatic digestion of genomic DNA (2-10 μg) to 2′-deoxynucleosides, followed by offline HPLC enrichment of MeC, hmC, fC, and caC and quantitation by capillary HPLC-ESI-MS/MS. Absolute quantitation is achieved by isotope dilution with the corresponding isotopically labeled internal standards ($^{13}C_{10}{}^{15}N_2$-MeC, D_3-hmC, $^{13}C_{10}{}^{15}N_2$-fC, $^{13}C_{10}{}^{15}N_2$-caC), which were synthesized in our laboratory as described in the methods section.[20] To improve HPLC retention and HPLC-ESI$^+$-MS/MS sensitivity for fC, it was derivatized with a biotin-containing aldehyde reactive probe.[14] HPLC-ESI-MS/MS detection of MeC, hmC, and fC was conducted by monitoring a neutral loss of deoxyribose (116 mass units) from each nucleoside (**Figure 4.2**), while a loss of deoxyribose and carbon dioxide was monitored for caC (**Figure 4.2**). Quantitative HPLC-ESI$^+$-MS/MS methodology was validated by analyzing synthetic DNA spiked with known concentrations of each analyte (**Figure 4.3**). The method's LOD was 3.1 fmol for MeC, 6.2 fmol for hmC, 1.2 fmol for fC, and 5 fmol for caC, while the corresponding LOQ values for MeC, hmC, fC, and caC were 6.2 fmol, 9.3 fmol, 4.9 fmol, and 8 fmol respectively, in 1 μg of genomic DNA. Excellent accuracy and precision were achieved for analyzing known amounts of MeC, hmC, and fC spiked into blank DNA. The sensitivity of our

analytical methodology was sufficient to quantify global amounts of MeC, hmC and fC in 1-2 μg of genomic DNA (**Figures 4.2** and **4.4**).

Figure 4.2. Representative capillary HPLC-ESI-MS/MS traces for accurate quantification of MeC, hmC, fC, and caC in mouse DNA.

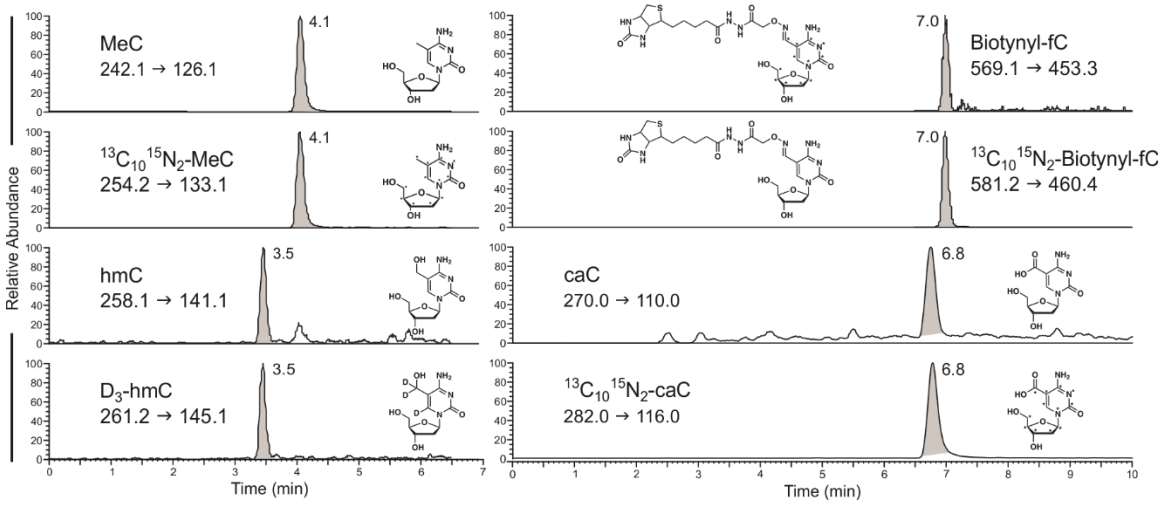
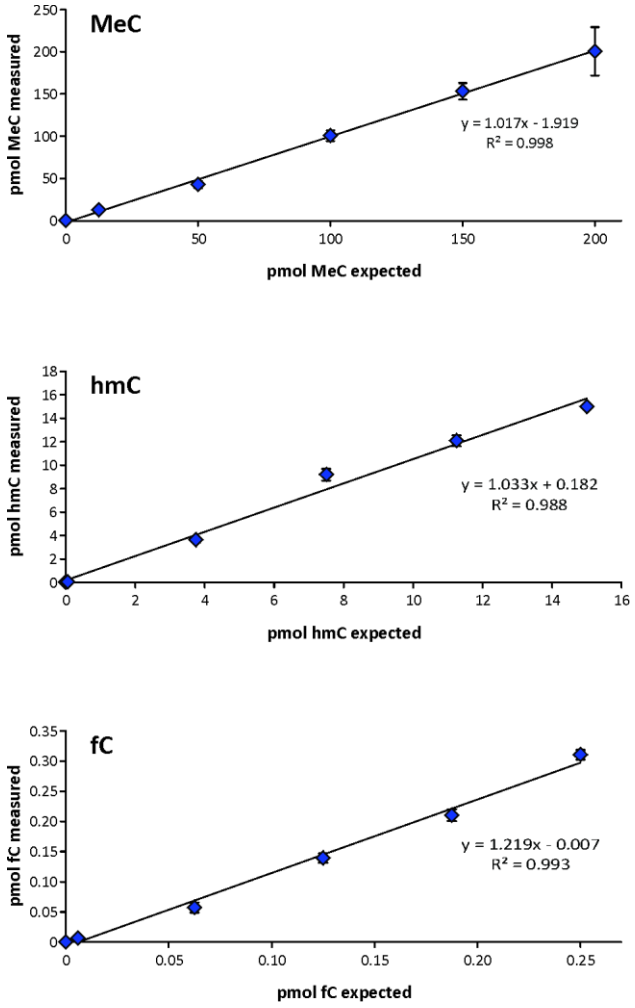


Figure 4.3. HPLC-ESI-MS/MS validation curves for MeC, hmC, and fC. Fixed amounts of isotopically labeled internal standards (1 pmol $^{13}C_{10}{}^{15}N_2$-MeC, 900 fmol d_3-hmC, and 500 fmol $^{13}C_{10}{}^{15}N_2$-fC) were spiked into calf thymus DNA and along with increasing amounts of the corresponding unlabeled nucleosides. Samples were processed as described above and subjected to HPLC-ESI$^+$-MS/MS analysis.



DNA methylation and hydroxymethylation changes in lung DNA of A/J mice exposed to cigarette smoke

Mass spectrometry based approach was used to quantify global cytosine methylation, hydroxymethylation, and formylation in lung DNA of A/J mice following whole-body exposure to environmental cigarette smoke (ECS) for 10 weeks, starting at birth, with or without oral co-administration of aspirin. These tissues were available form an earlier study by La Maestra et al.[140] These authors previously showed that while no histopathological change were apparent in the lungs of these mice, ECS caused a considerable downregulation of pulmonary microRNAs affecting both adaptive mechanisms and disease-related pathways.[291] Most notably, aspirin modulated the expression of microRNAs with a variety of functions including regulation of cyclooxygenases and inflammation.[291] We found that global levels of cytosine methylation, hydroxymethylation, and formylation remained relatively stable upon exposure to ECS or aspirin treatment (**Figure 4.4**).

To probe for site specific methylation and hydroxymethylation changes in DNA induced by ECS exposure, we conducted reduced representation bisulfite sequencing (RRBS) and oxidative-RRBS (oxoRRBS).[205] DNA isolated form lung tissues of female A/J mice (ECS and control groups, N = 4 per group) was prepared for sequencing using NuGEN Ovation RRBS Methyl-Seq with TrueMethyl oxBS modules.[205] The libraries were sequenced to a depth of 15-20 million reads. The amount of methylation at a covered cytosine could be inferred directly from the oxoRRBS data using MethPipe.[195] To examine

the changes in site-specific levels of hydroxymethylation, we first calculated the amounts of hmC present at each site from the RRBS and oxoRRBS data (See **Section 1.4.8** and **Figure 1.8** of Chapter I of this thesis for more information).[292] From these data, the sites with changes in methylation or hydroxymethylation greater than 5%, having more than 200 reads, and showing statistically significant changes in methylation[293] or hydroxymethylation were mapped to the nearest gene.

Site-specific methylation changes in ECS exposed mice ranged from increases of 30% to decreases of 36%. For each of the differentially methylated site (1537 sites), a methylation ratio was calculated in both the ECS treated and control samples. Methylation ratios are calculated by dividing the number of methylated reads by the coverage of the site. The methylation ratios are shown in a methylation heat map (Figure 4.4) normalized by z-score with hierarchal clustering. Hydroxymethylation changes calculated from the RRBS and oxoRRBS data ranged from increases of 56% to decreases of 47%. For each site with differential hydroxymethylation (N = 5524) a hydroxymethylation ratio was calculated. The ratios are shown in a heat map (**Figure 4.5**) normalized by z-score.

CpG sites showing increases and decreases in methylation were analyzed separately using Ingenuity Pathway Analysis (IPA) to identify any enrichments of associated functions and diseases. From this analyses, we found that genes with increased methylation (N = 193) were enriched for the formation of tumors, carcinomas, and development ($p < 1 \times 10^{-6}$, **Table 4.3**). The genes having a corresponding decrease in methylation (N = 357)

were analyzed in a similar fashion and found to also be associated with cancer and tumor development (**Table 4.4**).

Genes showing more than a 5% increase in hydroxymethylation (N = 1762) were also analyzed by IPA in the same manner to identify enrichment of the associated diseases and functions. The diseases and functions show a very high enrichment for genes associated with cancer, in particular, particularly adenocarcinoma (**Table 4.5**). Decreased levels of hmC was also observed in numerous genes (1350 genes), for which the enriched diseases and functions were determined. From this analysis, genes showing a decrease in hmC are also enriched in cancer associated genes (**Table 4.6**).

Figure 4.4. Global amounts of MeC, hmC, and fC in lung DNA of mice exposed to cigarette smoke for 10 weeks with or without aspirin co-treatment.

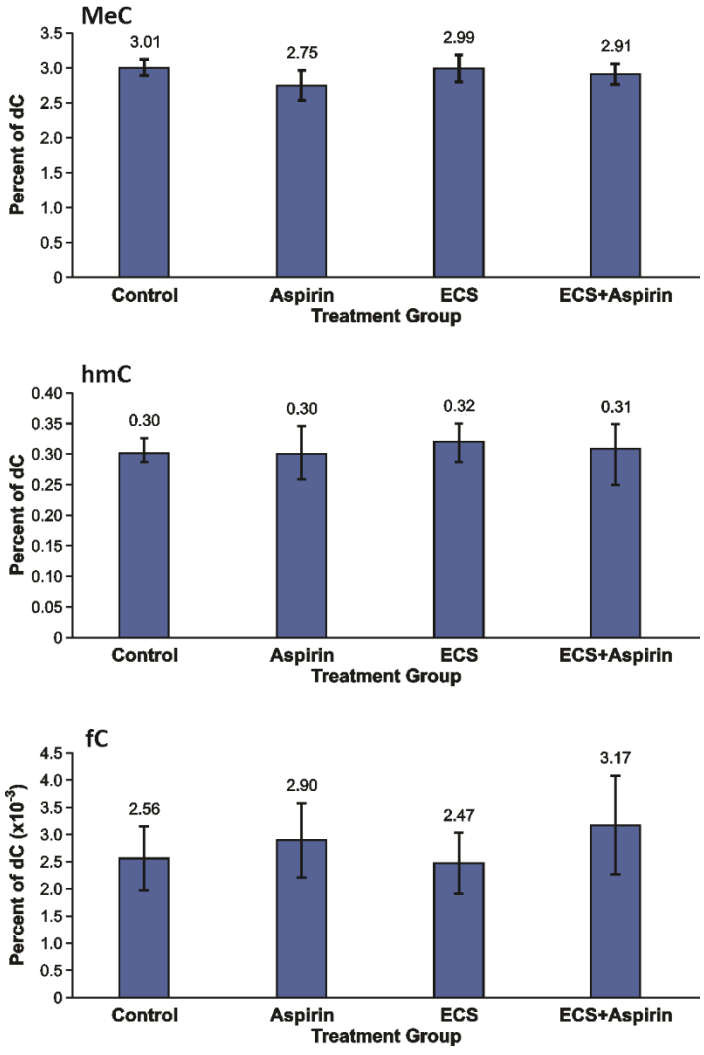


Figure 4.5. Site-specific cytosine methylation and hydroxymethylation levels in lung DNA of female A/J mice exposed to ETC for 10 weeks. Sites exhibiting > 5% change are included. **A**. A heatmap of the methylation ratio at specific sites in control and ECS exposed mice (>1500 sites). Data is expressed as methylation ratios (0-1). **B**. A heatmap of the hydroxymethylation ratio at specific sites in control and ECS exposed mice (>5500 sites). Data is expressed as normalized z-scores.

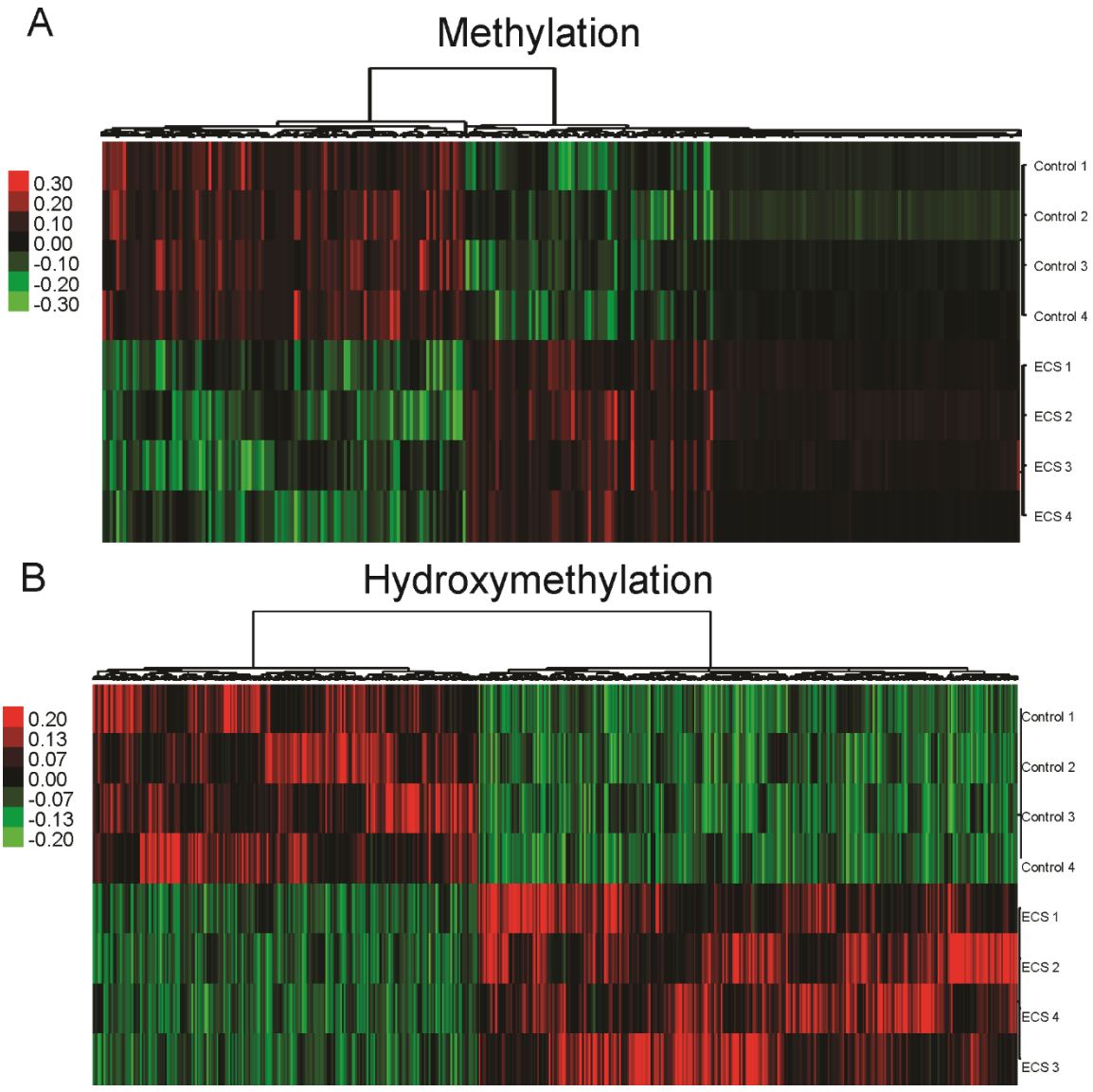


Table 4.3. Biological functions and diseases association of genes exhibiting increased cytosine methylation in the lungs of female A/J mice exposed to cigarette smoke for 10 weeks. Top 100 genes were analyzed using IPA.

Function Annotation	p-value
Liver tumor	1.35E-06
Development of adenocarcinoma	2.68E-06
Abnormality of endometrium	3.14E-06
Organismal death	3.25E-06
Endometrial carcinoma	4.97E-06
Uterine carcinoma	5.26E-06
Development of genital tumor	5.94E-06
Female genital tract adenocarcinoma	6.12E-06
Development of head	6.29E-06
Endometrioid carcinoma	6.57E-06

Table 4.4. Disease and functional enrichment of genes hypomethylated in mouse lung DNA following cigarette smoke exposure for 10 weeks. Top 100 genes by magnitude of MeC decrease were analyzed using IPA.

Function Annotation	p-value
Pelvic cancer	6.77E-09
Pelvic tumor	1.25E-08
Cancer of secretory structure	2.44E-08
Genital tract cancer	2.8E-08
Genital tumor	4.89E-08
Male genital neoplasm	6E-08
Prostatic tumor	6.46E-08
Malignant genitourinary solid tumor	8.56E-08
Malignant neoplasm of male genital organ	9.03E-08
Prostatic carcinoma	1.54E-07

Table 4.5. Disease and functional enrichment of genes exhibiting increased cytosine hydroxymethylation (hmC) in the lung following exposure of A/J mice to cigarette smoke for 10 weeks. The top 100 genes by magnitude of hmC increase were analyzed using IPA.

Function Annotation	p-value
Large intestine carcinoma	6.14E-36
Gastrointestinal adenocarcinoma	9E-36
Large intestine adenocarcinoma	1.38E-35
Gastrointestinal carcinoma	1.6E-35
Abdominal adenocarcinoma	2.52E-35
Liver lesion	1.55E-33
Non-melanoma solid tumor	2.72E-33
Non-hematological solid tumor	8.99E-33
Liver tumor	1.35E-32
Gastrointestinal tumor	1.42E-32

Table 4.6. Disease and functional enrichment for genes with decreased hmC following exposure of A/J mice to cigarette smoke for 10 weeks. The top 100 genes by magnitude of hmC decrease were analyzed using IPA.

Function Annotation	p-value
Gastrointestinal adenocarcinoma	7.23E-36
Large intestine adenocarcinoma	4.93E-35
Large intestine carcinoma	5.15E-35
Gastrointestinal carcinoma	1.17E-33
Intestinal tumor	1.02E-32
Large intestine neoplasm	1.45E-32
Gastrointestinal tumor	1.69E-32
Adenocarcinoma	2.47E-32
Malignant neoplasm of large intestine	4.75E-32
Abdominal adenocarcinoma	9.45E-32

Early epigenetic changes in the lung of A/J mice treated with tobacco carcinogen NNK and inflammatory agent LPS

While our RRBS data presented above provided evidence of epigenetic changes in lung DNA of mice exposed to cigarette smoke (**Figure 4.5**), it not clear which components of tobacco smoke are responsible. Cigarette smoke contains a complex mixture of chemicals (over 7,000), including tobacco specific nitrosamine NNK (118 ng/cigarette)[294] and bacterial endotoxin lipopolysaccharide (LPS) (2,400 ng/pack).[105] LPS is a major component of the outer membrane of Gram-negative bacteria. Upon inhalation, LPS induces immune response and inflammation by binding the CD14/TLR4/MD2 receptor complex in monocytes, dendritic cells, macrophages and B cells, promoting the secretion of pro-inflammatory cytokines, nitric oxide, and eicosanoids.[295]

In order to identify cigarette smoke components responsible for epigenetic effects of ECS (**Figure 4.5**), A/J mice were treated with NNK, LPS, or both NNK and LPS for 1-15 days. HPLC-ESI-MS/MS was used to quantify global MeC, hmC, fC, and caC in target (lung) and non-target tissues (kidney, brain). The amounts of MeC and hmC observed in lung tissue of A/J mice (3-3.5% of total cytosines being methylated and 0.1-0.12% total cytosine being hydroxymethylated) were comparable to the levels of MeC and hmC previously reported in lung tissue of C57BL/6N mice (4.2% of total cytosines for MeC and 0.15% for hmC).[30] We found that the global genomic levels of MeC and hmC in mouse lung were essentially unchanged over the two week treatment with NNK (**Figure 4.6**). However, global hmC concentrations were significantly lower in lung tissues of mice

treated with LPS (8.3 µg twice in week 1 and 4.15 µg once in week 2, $p < 0.05$, see **Figure 4.6**) or LPS/NNK in combination (0.13 ± 0.012 % vs 0.071 ± 0.008 %, $p = 0.025$). No significant changes were observed in mouse brain and kidney DNA (**Figure 4.7**). Overall, these results provide initial evidence that NNK alone has little effect on epigenetic marks of DNA, while LPS treatment may reduce the global levels of hmC in the target tissue (lung) but not in non-target tissues such as brain and kidney.

Figure 4.6. NNK and LPS alter the global levels of epigenetic modifications, MeC and hmC, in the lung tissue of treated A/J mice. Female A/J mice (6 weeks of age) were treated with NNK, with LPS, or both for a length of 2 weeks. Data are expressed as percent of dC and represents mean values ± SD of at least three animals. Global levels of MeC and hmC in the lung tissue of mice treated for 2 weeks with NNK or in combination with LPS showed little change for the levels of MeC, however a significant decrease was observed in hmC in the combination of NNK and LPS and in the mice sacrificed a week after the final treatment.

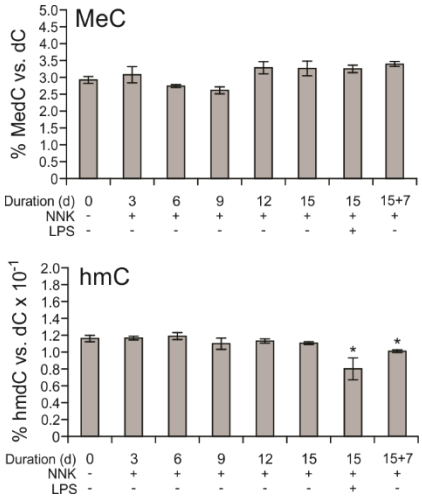
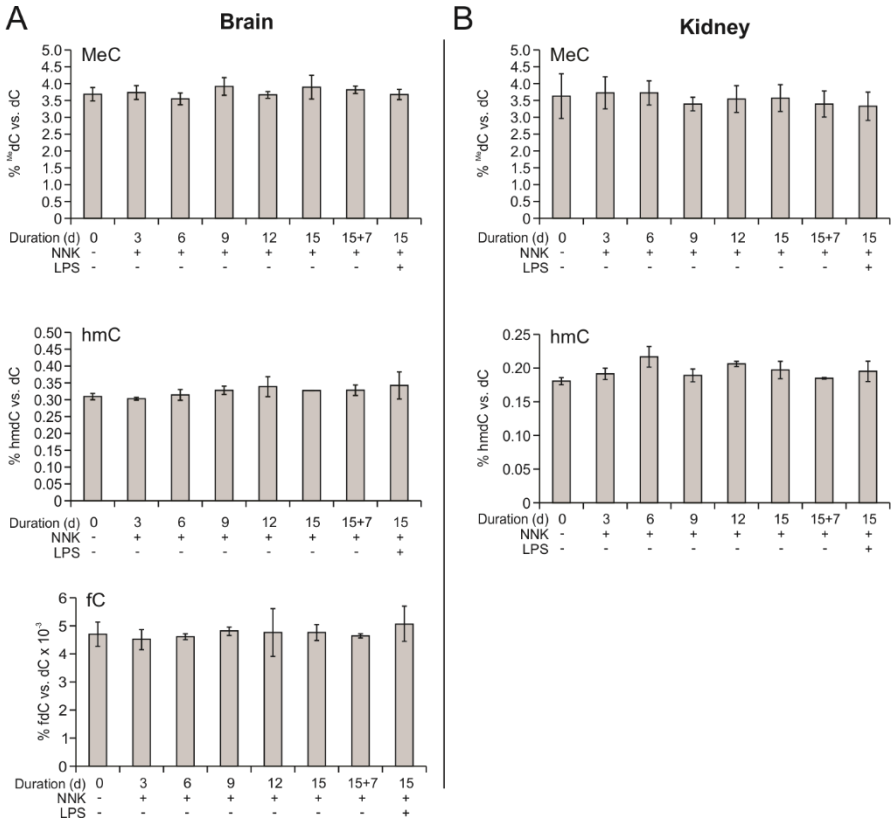


Figure 4.7. Early changes in global MeC and hmC levels in kidney and brain DNA of A/J mice treated with NNK and/or LPS. Female A/J mice (6 weeks of age) were treated intraperitoneally with NNK (25 mg/kg) every day on days 3 -9 and 12.5 mg/kg of NNK on days 10-14. For LPS treatments, 8.3 µg LPS was administered intranasally twice a week in the first week and 4.15 µg once during the second week. The error bars represent the average of N = 4 repeats for LPS-control and N = 3 for the rest of the groups.



Since epigenetic changes from the two week long treatment with tobacco carcinogens were relatively minor, a longer study was initiated. A/J mice were treated with NNK, LPS or NNK+LPS for six weeks, and the global amounts of epigenetic marks in lung DNA were determined by HPLC-ESI$^+$-MS/MS. As was the case in the short-term study (**Figure 4.7**), no significant changes in global MeC or hmC levels was observed for animals treated with NNK alone ($p = 0.66$ and 0.48, respectively, **Figure 4.8**). In contrast, HPLC-ESI$^+$-MS/MS analyses have revealed a small but significant increase in global amounts of MeC in lung DNA of mice treated with LPS alone (3.32 ± 0.12 % vs 3.52 ± 0.09 %, $p < 0.05$) or NNK/LPS in combination (3.32 ± 0.12 % vs 3.56 ± 0.05 %, $p < 0.05$) (**Figure 4.8**). Furthermore, global hmC levels were significantly decreased in lung DNA of animals treated with LPS or with NNK/LPS (0.12 ± 0.006 % and 0.13 ± 0.009 % C, respectively as compared to 0.19 ± 0.007 % in control animals, $p < 0.01$) (**Figure 4.8**). An opposite trend was observed for fC, which was elevated in LPS and NNK/LPS treatment groups 0.0058 ± 0.0015% $p = 0.05$ and 0.0058 ± 0.0003% C $p < 0.01$ respectively as compared to 0.0026 ± 0.0007% in control, **Figure 4.8**). Overall, our quantitative HPLC-ESI$^+$-MS/MS results for epigenetic DNA marks indicate that LPS-induced inflammation causes significant global hydroxymethylation changes in the lung.

A different trend was observed in brain tissues of the same animals. In this case, global cytosine methylation was increased in groups treated with NNK, LPS, or NNK/LPS, while hmC remained unchanged **(Figure 4.9**). Global fC levels in the brain showed an increase for groups treated with NNK, LPS, but not NNK/LPS (**Figure 4.9**). Due to the

relatively high abundance of oxidized forms of MeC in the brain, we were able to detect and quantify caC in this tissue (0.0012% ± 0.0003%), with amounts showing no change across treatment groups (**Figure 4.9**). These results indicate that treatment with LPS and NNK induces tissue dependent epigenetic changes, with most pronounced effects detected in the lung.

Figure 4.8. NNK and LPS alter the global levels of epigenetic modifications, MeC, hmC, and fC in the lung tissue of treated A/J mice. Female A/J mice (6 weeks of age) were treated intraperitoneally with NNK, intranasally with LPS, or both in combination for a length of 6 weeks. Data are expressed as percent of dC and represents mean values ± SD of at least three animals. LPS induced lung tissue inflammation drives changes in the global levels of MeC, hmC, and fC whether alone, or in co-exposure with NNK.

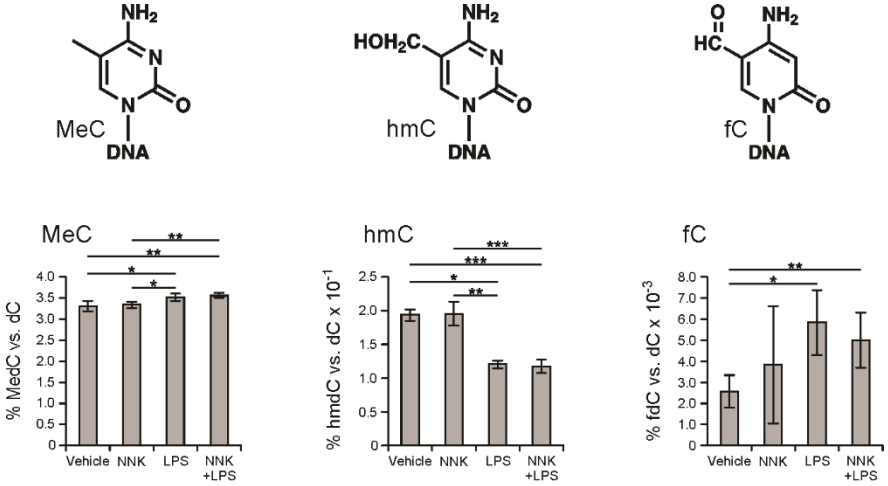
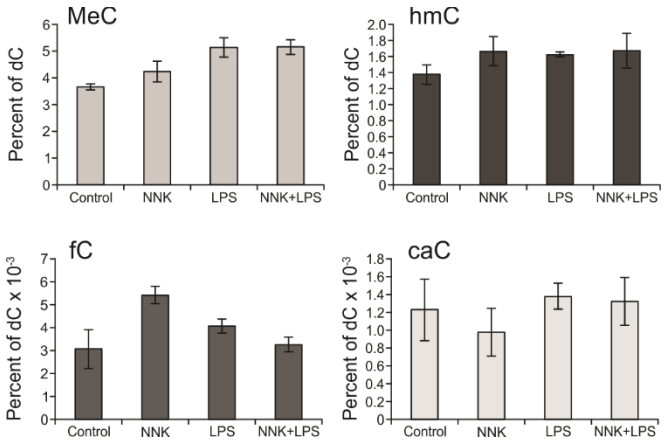


Figure 4.9. Global changes in genomic MeC, hmC, fC, and caC levels in the brain of A/J mice chronically treated with NNK, LPS, or both NNK and LPS for 5 weeks. NNK (50 mg/kg) was given IP twice during weeks 1 and 3, and once during week 5. LPS (4 µg/mouse) was administered intranasally twice during week 1, and once a week in weeks 3, 4, and 5. The error bars represent the average of N = 4 repeats.

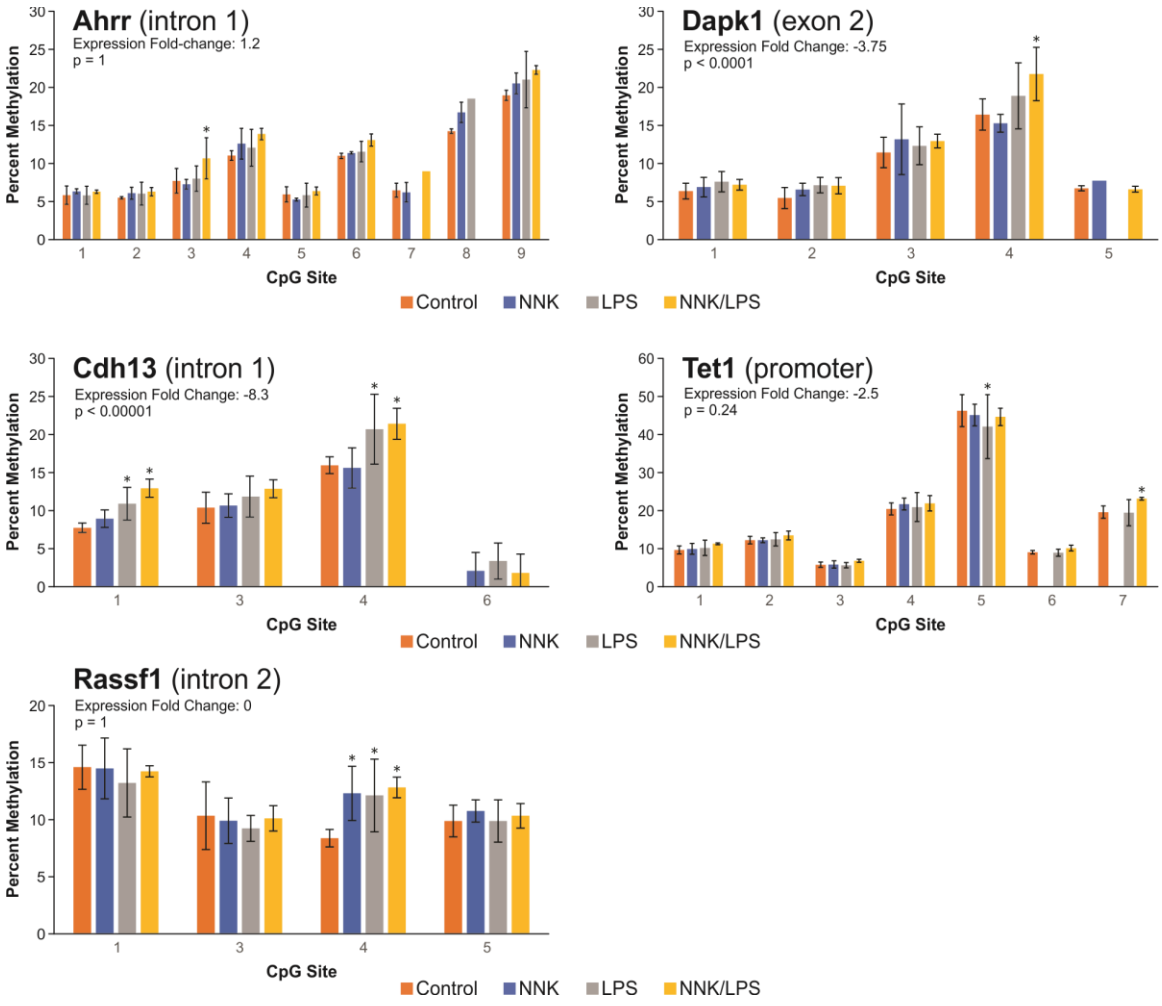


DNA methylation levels at specific genes in DNA from 6-week chronic mouse treatment

To detected site-specific methylation changes following chronic exposure to NNK and LPS, pyrosequencing study of five genes (Ahrr, DAPK1, CDH13, Tet1, and Rassf1) was conducted. These genes were selected for further study due to their association with cigarette smoke studies in humans.[296] These experiments revealed a range of methylation at specific CpGs in each gene, CDH13 2-23%, Ahrr 5-24%, DAPK1 5-23%, Tet1 5-46%, and Rassf1 9-14% (**Figure 4.10**). The methylation level of the Ahrr gene was measured at 9 CpG sites. In the Ahrr gene, small increases (0.3-2.5%) were observed for NNK and LPS treatment alone. In the combined treatment of NNK and LPS an increase in methylation was observed for CpGs 3 (+3%, $p < 0.05$), 4 (+2.8%), 6 (+2.1%), and 9 (+3.3%). Pyrosequencing of the DAPK1 gene provided methylation levels for 5 CpG sites. DAPK1 methylation was increased at each CpG and for each condition except in the NNK treatment at CpG 4 which showed a slightly decreased methylation level. In the combination treatment of NNK and LPS, the methylation level of each CpG site in DAPK1 increased 0.8%, 1.6%, 1.5%, and 5.3% ($p < 0.01$) for sites 1, 2, 3, and 4 respectively and decreased 0.1% at site 5. In the CDH13 gene, site specific methylation was increased at CpGs 1, 3, and 4 in both the LPS alone and in the combination of NNK and LPS. The increases at these CpG sites for LPS treatment were 3.2% ($p < 0.05$), 1.5%, and 4.8% ($p < 0.01$) and for treatment with NNK and LPS were 5.2% ($p < 0.01$), 2.5%, and 5.5% ($p < 0.01$) respectively. Pyrosequencing detected site specific methylation changes in the Tet1

promoter with NNK treatment causing a +0.3%, 0%, +0.1%, +1.3%, and -1.2% for CpGs 1, 2, 3, 4, and 5 respectively while LPS treatment caused +0.6%, +0.2%, -0.1%, +0.5%, and -4.2% ($p = 0.01$) for those same sites. Combined NNK + LPS treatment caused increased methylation at all CpGs measured (except CpG 5), with methylation changes of +1.6%, +1.2%, +1.0%, +1.5%, -1.6%, +1.1%, and +3.5% ($p < 0.05$) at each of those sites. Site-specific methylation of the Rassf1 gene was measured at 4 CpG sites by pyrosequencing. Cytosine methylation of Rassf1 was decreased at CpGs 1 and 2 in all conditions. At CpG 3 in Rassf1, methylation levels were increased 3.9% ($p < 0.05$), 3.7% ($p < 0.05$), and 4.5% ($p = 0.01$) and at CpG 4 the methylation level increased 0.9% and 0.5% in NNK and the combination of NNK and LPS treatment.

Figure 4.10. Methylation status of specific CpG sites within promoter regions of Ahrr, DAPK1, CDH13, Tet1, and Rassf1 genes in lung DNA of mice treated for 6 weeks with vehicle, NNK, LPS, and NNK/LPS determined by bisulfite pyrosequencing. * indicates $p < 0.05$ vs. control.



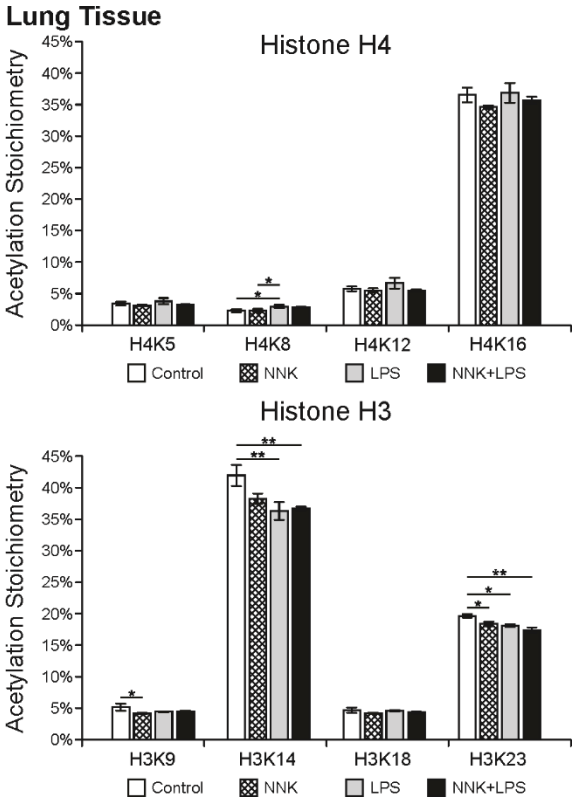
Histone acetylation changes

Histone acetylation is a key regulatory mechanism that maintains epigenetic memory in eukaryotic cells.[297] Site-specific acetylation of histones H3 and H4 strongly correlates with the global gene expression profiles and transcriptional activation activities.[298-299] On the other hand, dysregulation of histone acetylation contributes to tumorigenesis and disease progression in lung cancer and other types of carcinomas.[298, 300] To determine whether NNK and LPS treatments induce global changes in histone acetylation, we employed a recently developed chemical proteomics strategy that enables site-specific identification of the absolute lysine acetylation stoichiometries.[290]

Histone proteins were extracted from lung tissues from A/J mice treated for 6 weeks, labeled, and analyzed by nano HPLC-ESI$^+$-MS/MS as described previously.[289] The data were analyzed by in-house developed scripts to calculate site-specific acetylation stoichiometries.[290] Our data shown in **Figure 4.11** demonstrate that NNK and LPS treatments led to the overall down-regulation of lysine acetylation abundance on histone H3 and H4. More specifically, acetylation on histone H3K14 and H3K23 significantly decreased upon the treatment with NNK, LPS alone, or NNK/LPS, while the treatments did not significantly affect H3K18 acetylation (**Figure 4.11**). Histone H4 acetylation was also down-regulated by NNK and LPS treatment, but to a lesser extent than H3. NNK treatment alone and the combination of NNK/LPS treatment significantly downregulated histone H4K16 acetylation, while the treatments did not strongly decrease other histone H4 N-terminal acetylation (**Figure 4.11**). It is worth noting that acetylation of H3K14, H3K23

and H4K16 are transcriptional activation marks that are strongly associated with active gene expression.[299, 301-304]

Figure 4.11. Histone lysine acetylation stoichiometry changes upon treatment of A/J mice with NNK, LPS, and coexposure to NNK and LPS for 6 weeks. Data is expressed as percent acetylated and represents mean values ± SD of at least 3 animals. Statistical significance was evaluated between treated and control samples using a two-sided Student's t-test (* $p < 0.05$, ** $p < 0.01$). Histone acetylation stoichiometries were assessed to determine global gene activation by deacetylation due to exposure. Measured acetylation sites include lysines 5, 8, 12, and 16 on histone H4 and lysines 9, 14, 18, and 23 on Histone H3.

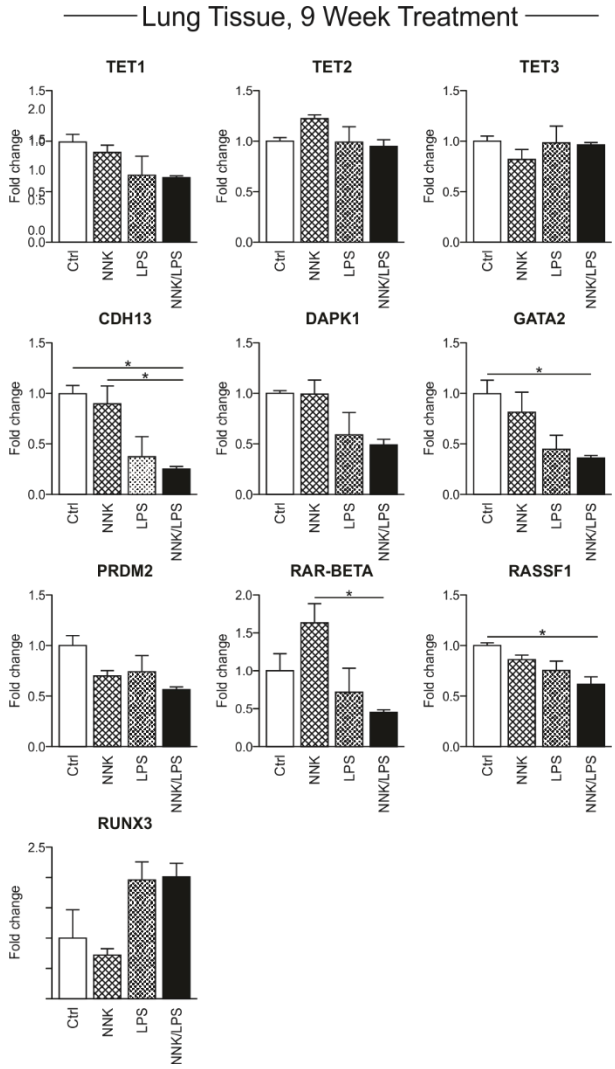


Gene expression levels in lung tissues of NNK/LPS treated mice

To determine whether LPS- and NNK-mediated changes in DNA and histone epigenetic marks lead to altered gene expression levels, we examined the expression levels of genes including *Tet1*, *Tet2*, *Tet3*, *Dapk1*, *Gata2*, *Cdh13*, *Prdm2*, and *Rassf1*, and transcription factor *Runx3* using qRT-PCR (**Figure 4.12**). We found that *Tet1* expression was decreased in all treatment groups, while *Tet* isoforms 2 and 3 were unaffected. NNK/LPS treated animals showed more pronounced changes in Tet1 mRNA as compared to NNK only group ($p < 0.05$, see **Figure 4.12**). These results suggest that LPS-induced inflammation is associated with reduced levels of expression for the *Tet1* gene, which is consistent with our observation of decreased levels of hmC in genomic DNA (**Figures 4.4 and 4.6**). In addition, gene expression levels of tumor suppressor genes *Dapk1*, *Gata2*, *Cdh13*, *Prdm2*, and *Rassf1* showed significant changes in the lungs of treated mice. The mRNA level of death-associate protein kinase 1 (*Dapk1*) showed a 2-fold decrease in both LPS and treated groups but was unchanged in NNK only group (**Figure 4.12**). Similarly, the expression levels of cellular adhesion protein cadherin 13 gene (*Cdh13*) decreased 3-fold in both groups treated with LPS, but not in NNK only group (**Figure 4.12**). Expression levels of the retinoic acid receptor, beta (*Rar-β*) increased upon treatment with NNK but decreased in LPS alone and NNK/LPS groups (**Figure 4.12**). For the zinc-finger transcription factor *Gata2*, mRNA levels were reduced in each treatment group (**Figure 4.12**) and was most significant in the NNK/LPS group, $p < 0.05$. The mRNA levels for PR Domain containing protein 2 (*Prdm2*) decreased slightly with treatment relative to control

(**Figure 4.12**). A small decrease in Ras association domain family member 1 (*Rassf1*) was observed in all treatment groups ($p < 0.05$) (**Figure 4.12**). In contrast, the expression of Runt related transcription factor 3 (*Runx3*) was elevated in both LPS treated groups but was unchanged in NNK only group. Overall, significant changes in expression levels of many critical genes were observed in the lung of mice treated with LPS to induce inflammation, but the effects of NNK alone on gene expression were relatively small.

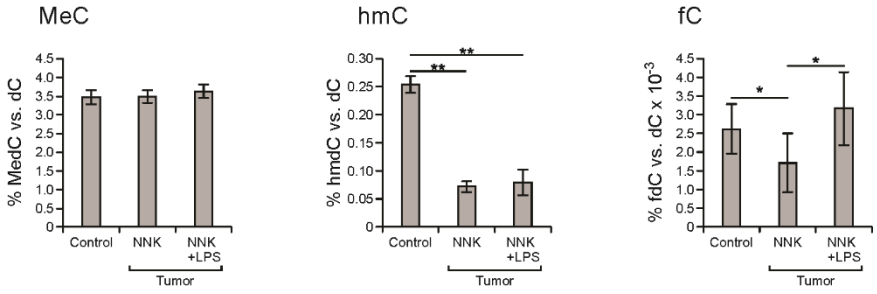
Figure 4.12. Gene expression changes in the lung tissues of A/J mice treated with treated with a single dose of NNK (100 mg/kg, IP) in week 1, weekly LPS (5 µg/mouse in 50 µL PBS, intranasally under isoflurane anesthesia) starting week 2, both NNK and LPS or physiological saline only (control) for a total of 9 weeks. Data are calculated using the ΔΔCt method ± SD with three biological and three technical replicates. Genes were selected based on association with lung cancer.



Epigenetic changes in NNK/LPS induced pulmonary adenocarcinoma

To determine whether tobacco carcinogen-induced epigenetic changes persist throughout lung cancer development, MeC, hmC, and fC were quantified in mouse pulmonary adenocarcinomas induced by NNK/LPS treatment. Tumors (≥ 100) isolated from the lungs of 4-5 A/J mice that had been treated with NNK alone or NNK/LPS for 27 weeks were excised and pooled.[141] Genomic DNA was extracted from pooled tumor samples and analyzed for global levels of MeC, hmC, and fC by isotope dilution HPLC-ESI^+-MS/MS. The values were compared to non-tumor DNA from lung tissue of mice of the same age. We found that while global levels of MeC remained relatively stable across treatment groups (3.5 ± 0.2 - 3.6 ± 0.2 % of total Cs, $p = 0.15$, **Figure 4.13**), both hmC and fC levels were altered in tumors. Cytosine hydroxymethylation decreased significantly in DNA extracted from NNK- and NNK/LPS-induced tumors (0.07 and 0.08 %, respectively, as compared to 0.25 ± 0.015% of total Cs in normal lung tissue ($p < 0.0001$ and 0.0001, respectively, see **Figure 4.13**). Global amounts of fC showed a small decrease in NNK-induced tumors (0.0026 ± 0.0007 vs 0.0017 ± 0.0008 % of total Cs, $p = 0.027$) and were increased in tumors induced by NNK/LPS treatment (0.0026 vs 0.0032 ± 0.001) although this difference was not statistically significant (**Figure 4.13**). Overall, global amounts of MeC, hmC, and fC in lung tumors exhibited the same overall trend as lung tissues prior to tumor formation (compare **Figures 4.6** and **4.13**).

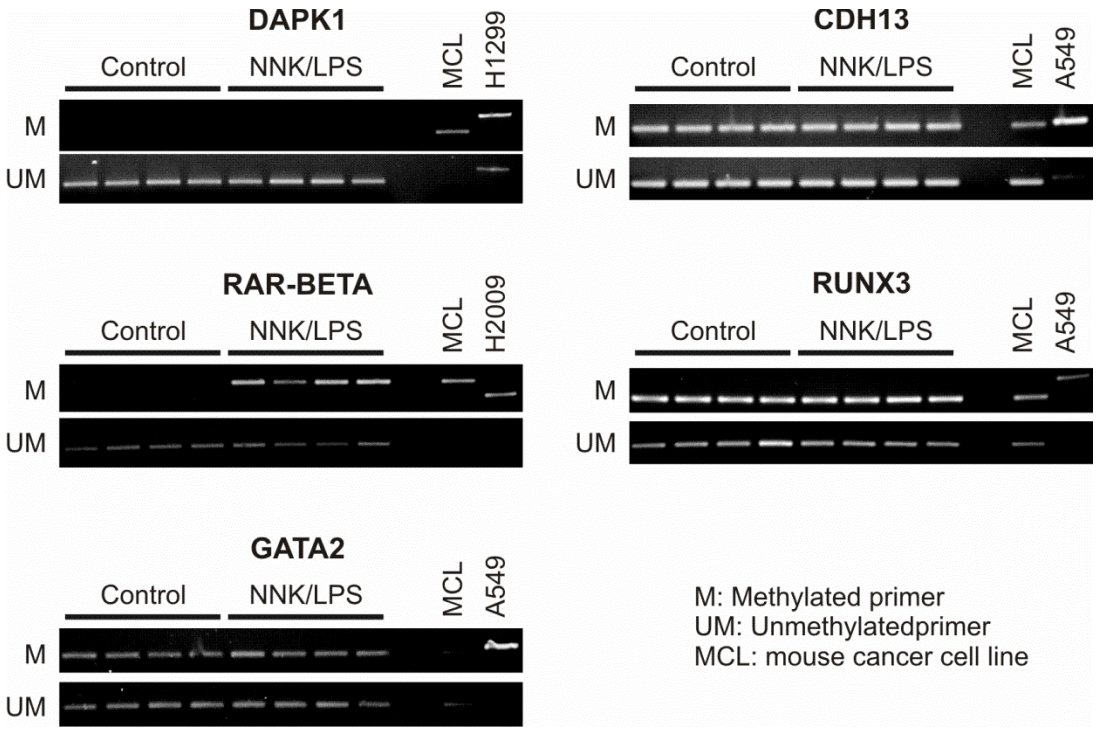
Figure 4.13. NNK and LPS alter the global levels of epigenetic modifications, MeC, hmC, and fC in the lung tumors and normal control tissue of A/J mice. Female A/J mice (6 weeks of age) were treated intraperitoneally with NNK, intranasally with LPS, or in combination for a length of 27 weeks. Data are expressed as percent of dC and represents mean values ± SD of at least three animals. 27-week exposure to NNK or the combination of NNK and LPS to generate lung tumors show global levels of MeC, hmC, and fC are deregulated.



Promoter methylation of tumor suppressor proteins in tumors of A/J mice

Methylation specific PCR analysis was conducted to determine whether global methylation and hydroxymethylation changes in tumors are accompanied by altered promoter methylation and hydroxymethylation of tumor suppressor genes. We found that the *Dapk1* promoter was unmethylated in all samples, while the methylation of *Cdh13*, *Runx3*, and *Gata2* promoters was unchanged between tumors and controls. However, *Rar-β* promoter was methylated only in treated group but not in control lung (**Figure 4.14**).

Figure 4.14. Methylation specific PCR results from DNA isolated from lung tumors of A/J mice treated for 44 weeks with a single dose of NNK (100 mg/kg, IP) in week 1 and biweekly dosing of LPS (2 µg/mouse in 50 µL PBS, intranasally under isoflurane anesthesia) beginning in week 2.



Gene expression levels in lung tumors of NNK/LPS treated mice

To understand whether initial inflammation-induced changes in gene expression in the lung persist upon tumor formation, mRNA expression analyses were repeated in lung tumors of A/J mice formed 22 and 44 weeks post treatment with NNK, LPS, or with NNK/LPS. We observed pronounced gene expression changes in tumors (22-week tumors – **Figure 4.15**, and 44-week tumors - **Figure 4.16**). These results support the notion that inflammation mediated changes in gene expression contribute to carcinogenesis in the A/J mouse model of lung cancer.

Figure 4.15. Gene levels of TET proteins and tumor suppressor genes changes in lung tumors of A/J mice. Mice were treated with treated with a single dose of NNK (100 mg/kg, IP) in week 1, weekly LPS (5 µg/mouse in 50 µL PBS, intranasally under isoflurane anesthesia) starting week 2, both NNK and LPS or physiological saline only (control). Under these conditions, lung tumors were formed for both the NNK and NNK with LPS treatment groups for 22 weeks. Data were calculated from qRT-PCR using the ΔΔCt method ± SD with three biological and three technical replicates.

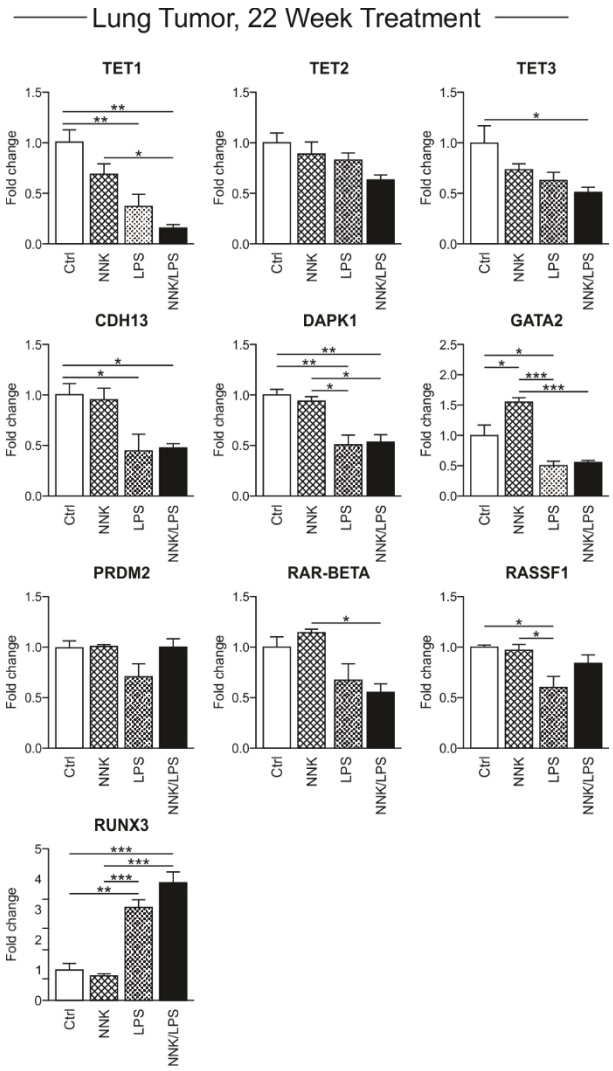
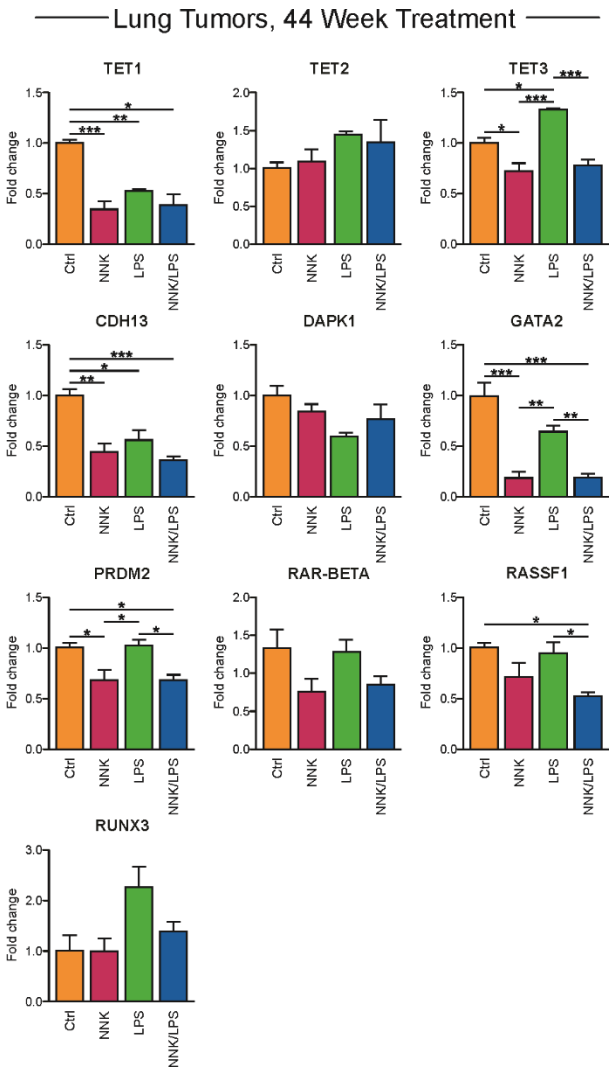


Figure 4.16. Gene expression changes in the lung tumors of A/J mice treated with a single dose of NNK (100 mg/kg, IP) in week 1, biweekly LPS (2 µg/dose in 50 µL PBS, intranasally under isoflurane anesthesia) starting week 2, or both NNK/LPS for a duration of 44 weeks as compared to normal lung tissue control. Data are calculated using the ΔΔCt method ± SD with three biological and three technical replicates.



4.4 Discussion

Lung cancer kills over one million people each year worldwide and is the leading cause of cancer death in the United States.[305] With a 5 year survival rate of 5-54% depending on stage of disease, new prevention and treatment strategies for lung cancer are urgently needed.[305] The primary risk factor for lung cancer development is smoking. Although the prevalence of smoking among adults in the United States has been decreasing, 16.8% of Americans continue to smoke.[306] Cigarette smoke contains over 60 known carcinogens including tobacco specific nitrosamine NNK.[307] Following metabolic activation to reactive metabolites, NNK induces a range of promutagenic DNA adducts including O^6-methylguanine.[307-308] Error-prone polymerase bypass of DNA adducts leads to heritable mutations, which in turn contribute to lung cancer development.[307]

Tumor promotion and inflammatory stimuli, in combination with genotoxic stimuli, are increasingly recognized as important factors in smoking-induced carcinogenesis.[309] Tobacco smoke contains pro-inflammatory compounds such as lipopolysaccharide (LPS, 2,400 ng/pack).[105] When administered to laboratory mice, LPS induces inflammatory response mimicking chronic obstructive pulmonary disease (COPD), a major risk factor for lung cancer development in smokers.[283] While the exact mechanism of inflammation-driven lung tumor initiation and promotion is unknown, epigenetic changes are likely to play a critical role as other inflammation-driven cancers such as *H. pylori*-related gastric cancer,[310] dextran sulfate-induced colon cancer,[311] and

hepatitis C virus-induced hepatocellular carcinoma[312] are characterized by hypermethylation of tumor suppressor genes.

NSAIDs are commonly prescribed anti-inflammatory agents that have been shown to reduce the incidence of inflammation driven colorectal cancer.[285] Although the efficacy of NSAIDs in preventing lung cancer chemoprevention remains to be established,[286-288, 313] aspirin has been shown to reduce the number of NNK induced lung tumors in A/J mouse model of smoking-induced lung cancer.[314]

Irregular DNA methylation events, e.g. "epimutations", have been shown to collaborate with genetic changes to drive the cancer phenotype.[315-318] Epigenetic alterations such as aberrant cytosine methylation, abnormal histone acetylation, changes in micro RNA levels, and deregulated patterns of gene expression are increasingly recognized as key events in cancer etiology.[319-324] Tumor DNA sequencing as part of The Cancer Genome Atlas (TCGA) has revealed that components of epigenetic machinery are among the genes most frequently mutated in human cancers.[325] Many tumors are characterized by aberrant DNA methylation, decreased TET protein expression, and reduced levels of DNA hydroxymethylation.[326]

It is crucial that we learn more about the epigenetic pathways of tumorigenesis. Unlike somatic mutations, which are difficult to correct with current treatments, epigenetic changes can be reversed with appropriate therapeutic agents.[327] Epigenetic changes take place relatively soon following exposure and could be used as biomarkers of lung cancer susceptibility and to aid in early detection of lung tumors.[323, 328] Finally, by providing

individualized “fingerprints” of smoking-mediated gene expression changes, epigenetic markers can facilitate the development of personalized therapies for lung cancer,[329] Because MeC typically acts as a repressive mark, its oxidation to hmC may trigger gene reactivation. Therefore, diminished oxidation of MeC to hmC in tumors is likely contribute to inactivation of tumor suppressor genes, triggering malignant transformation.[330] However, the dynamics, mechanisms, and contributions of epigenetic changes to cancer etiology remain poorly understood.

Multiple studies have been conducted to determine the role of cytosine methylation in cancer. However, many of the earlier studies have been limited to investigating MeC at a single endpoint following tumor formation and did not examine early epigenetic changes. Furthermore, traditional sequencing methods did not distinguish between MeC and hmC, since both of them resist bisulfite conversion.[167, 204] Although gene-specific hypermethylation of *Dapk1*, *Rassf1a*, *Mgmt*, *Mlh1* and *p16* have been identified as epigenetic changes associated with lung tumor development, the timing of cell epigenetic transformations has not been identified.[108, 331-333] Cytosine hydroxymethylation, formylation, and carboxylation has been studied to a lesser extent than methylation, although global hmC has been consistently found to be depleted in tumors, correlating with a decrease in Tet gene expression.[334]

The main goal of the present study was to characterize epigenetic changes in the lung following cigarette smoke exposure or inflammatory stimuli/tobacco carcinogen treatment in an established A/J mouse model of smoking induced lung cancer.[141] A/J mice,

which were exposed to cigarette smoke starting 12 h after birth and continuing for 10 weeks,[140] were investigated for global and loci-specific changes in MeC and hmC using RRBS and oxoRRBS. The sites with changes in methylation and hydroxymethylation were assigned with the nearest gene and analyzed using IPA to identify any enrichment in diseases and functions associated with those genes. For each group of genes input into IPA, increased MeC, decreased MeC, increased hmC, and decreased hmC, the overwhelming enrichment focused on various cancers (**Tables 4.3-4.6**). This indicates that exposure to cigarette smoke induces epigenetic changes in site-specific levels of MeC and hmC at genes that are associated with carcinogenesis.

Changes in site specific methylation and hydroxymethylation observed by RRBS and oxoRRBS in mice exposed to cigarette smoke (**Figure 4.4**) prompted us to investigate the epigenetic effects of tobacco carcinogen NNK and inflammatory agent LPS. From these analyses, we report a measurable change in global DNA methylation and hydroxymethylation that triggers gene expression changes in the A/J mouse model of smoking induced lung cancer. We found that global levels of MeC, hmC, and fC were unchanged in NNK-treated animals, but were affected in animals treated with LPS only or a combination of NNK and LPS. A 33% change in global hmC levels was observed in mice treated with LPS for 15 days. However, changes in global MeC were not observable until chronic treatment with LPS, at which point LPS and NNK/LPS treatments increased the amount of MeC by 6% and 5% respectively. Similarly, fC was increased by 2-fold after the same chronic treatment with LPS and NNK/LPS. This suggests that measurable

changes in global hmC occurred after 15 days of exposure to inflammation, while measurable changes in MeC and fC occur between two and six weeks of exposure to LPS, which indicates that hmC may be an early indicator of epigenetic shift in our model.

Based on the site-specific changes in cytosine methylation and hydroxymethylation in cigarette smoke exposed mice and the changes observed in global levels of MeC and hmC following 6 week exposure to LPS, we investigated whether NNK, LPS, or both in combination altered site specific methylation by pyrosequencing. DAPK1, CDH13, and Rassf1 were selected based on changes observed in lung cancer patients,[110, 335-340] TET1 was selected based on the decreases observed in hmC at 6 weeks, and the AHRR gene was selected based on its association with cigarette smoking.[296, 341-342] In these genes, we observed an overall increase in CpG methylation due to the exposure to NNK, LPS, or NNK and LPS with both LPS and NNK plus LPS having the most pronounced changes. The increases were significant for LPS treatment in CDH13 (CpGs 1 and 4) and Rassf1 (CpG 3), and in NNK plus LPS treatment in Rassf1 (CpG 3), DAPK1 (CpG 4), CDH13 (CpGs 1 and 4), TET1 (CpG 7), and Rassf1 (CpG 3). These increases of methylation for tumor suppressor genes DAPK1, CDH13, Rassf1, and TET1 correspond to decreased expression observed in qRT-PCR of mice treated for 9 weeks. Our results showing increased methylation of key genes agree with previous work which showed that the expression of Rassf1a, CDH13 and DAPK1 is decreased in lung cancer due to increased gene methylation,[338, 343-344] In addition, DAPK1 and CDH13 have increased methylation after 6 weeks of exposure.

Histone lysine acetylation is generally associated with open chromatin and transcriptional activation.[297] Our data revealed globally decreased acetylation abundance on histone H3 and H4 N-terminal sites upon NNK and LPS treatment, suggesting a potential transcriptional repression effect under these chemical treatments. Hypoacetylation of histone H4K16 has been reported in lung cancer.[345] The phenomenon is strongly associated with the overexpression of histone deacetylases and the activity loss/mutations in histone acetyltransferases in lung cancer cells.[346] In addition, LPS treatment has been previously shown to induce histone deacetylase expression and reduce histone acetylation abundance.[347] Loss of histone acetylation, together with altered DNA methylation patterns, contribute to the accumulation of suppressive epigenetic marks and the decrease in expression of important tumor suppressor genes.

Gene expression analyses by qRT-PCR have revealed that the observed changes in global MeC, hmC, fC amounts coincide with a decrease in expression of *Tet1* and *Tet3* in NNK and LPS treated mice. These changes in gene expression support previous reports showing reduced *Tet* expression in tumors.[334] Although gene expression of *Tet1* and *Tet3* is decreased in lung tissue after 9 weeks of treatment with both NNK and NNK/LPS, global MeC and hmC in lung tissue were unchanged after 6 weeks in NNK treatment relative to control (**Figure 4.8**). In addition, *Tet1* gene expression levels in lung tissue after 9 weeks of treatment shows a 10% decrease from vehicle to NNK and an even greater (35%) decrease for combined NNK/LPS treatment group. These results suggest that NNK exposure and inflammation work synergistically at early stages of exposure to cigarette

smoke to reduce *Tet* expression in the lung. Significant changes in gene expression were observed after 9 weeks of treatment and persisted through 22 and 44 weeks and tumor development. Specifically, the genes for which expression is decreased by 20% or more at 9 weeks, such as *Cdh13*, *Prdm2*, *Gata2*, and *Rassf1* show significant decreases after tumor formation by both 22 and 44 weeks. *Gata2* is decreased 75% in NNK induced tumors relative to control compared to 20% in mouse lung tissue from mice treated for 9 weeks with NNK.

To investigate the mechanisms by which gene expression levels were affected, we explored promoter methylation for a subset of those genes. For *Rar-β*, we observed methylation of the promoter which corresponded to a decrease in gene expression. *Cdh13* promoter was methylated in both vehicle and NNK/LPS treated mice, making its role in regulation of gene expression unclear. These results suggest that gene promoter methylation is not the only mechanism responsible for controlling gene expression. It is possible that gene expression could be controlled by other regulating factors such as histone modifications, RNA processing, changes in N^6-methyladenosine increasing RNA lifetimes[348] or miRNA.[349]

Our results indicate global epigenetic changes in the lung precede the formation of tumors, and that epigenetic changes are mediated by inflammation. We also observe that accompanying the epigenetic changes were altered gene expression levels generally resulting in decreased expression. We propose that inflammation-mediated early epigenetic changes in the lung lead to profound changes in gene expression of tumor suppressor

proteins, contributing to lung cancer etiology. Future studies should focus on defining the timing and contributions of changes in hmC in lung carcinogenesis, and the mediators responsible.

Overall, our study using a well-established A/J mouse model for smoking induced lung cancer demonstrate that NNK and LPS treatments are sufficient to alter the global epigenetic landscape of cytosine methylation, cytosine hydroxymethylation, and histone acetylation, potentially predisposing pulmonary cells to the onset of tumorigenesis. In our future studies, oxoRRBS will be employed to characterize methylation changes with single nucleotide resolution in type II alveolar cells of laboratory mice exposed to cigarette smoke. Identification of aberrantly regulated histone acetylation sites upon NNK and LPS treatments will allow us to further investigate the mechanisms of chemical-induced alterations in the enzymatic activities of histone deacetylases and acetyltransferases that regulate these epigenetic marks during lung cancer progression. In addition, genome-wide ChIP-seq analysis targeting these histone acetylation sites will provide further details on the locus-specific dynamics of the epigenetic reprogramming induced by the chemical treatments. These data, oxoRRBS and ChIP-Seq, when combined with RNA-Seq analysis will allow us to determine specific cellular pathways that are transcriptionally suppressed and therefore contribute to the smoking mediated development of lung cancer.

V. EPIGENETIC EFFECTS OF INFLAMMATION IN TYPE II ALVEOLAR CELLS OF A/J MICE TREATED WITH LIPOPOLYSACCHARIDE

Adapted from:

Seiler, C. L.; Song, J.M.; Fernandez, J.; Abrahante, J.E.; Chen, Y.; Ren, Y.; Kassie, F.; and Tretyakova, N. Y.; Epigenetic Effects of Inflammation in Type II Alveolar Cells

Manuscript in preparation

This work was performed in collaboration with Dr. Jung Min Song, Jenna Fernandez, Dr. Juan E. Abrahante, Dr. Yue Chen, Yanan Ren, Dr. Fekadu Kassie, under the direction of Dr. Natalia Tretyakova. Dr. Jung Min Song and Jenna Fernandez performed the animal treatment under the direction of Dr. Fekadu Kassie. Dr. Yue Chen measured histone acetylation levels. Yanan Ren performed statistical calculations for the LC-MS/MS data. Christopher Seiler isolated DNA and mRNA from type II cells, carried out analysis of global epigenetic DNA marks, and conducted RNA-Seq with the University of Minnesota Genomics Center. Dr. Juan E. Abrahante provided bioinformatics support for RNA-Seq data. Christopher Seiler and Natalia Tretyakova wrote the manuscript.

5.1 Introduction

Lung cancer is a leading cause of cancer deaths world-wide and is expected to kill 154,050 Americans this year, with over 80% of cases directly attributable to smoking.[101] Cigarette smoke contains 60+ known carcinogens[104] and co-carcinogens, including bacterial endotoxin lipopolysaccharide (LPS) (2,400 ng/pack).[105] LPS induces an immune response in mammalian cells by interacting with the toll-like receptor 4 (TLR4).[295] This interaction is mediated by MD-2 and CD14, which are necessary for TLR4 recognition and shuttling LPS toward TLR4, respectively.[295] Through downstream signaling pathways, TLR4 activates NF-κB, which controls the expression of proinflammatory cytokines and immune related genes.[295]

Inflammation of the lung, which is observed in chronic obstructive pulmonary disease (COPD), increases the risk of developing lung cancer.[131, 350] The majority of smoking-induced lung cancers are defined as adenocarcinomas,[121] which can arise from the alveolar type II cells in the lung.[225] Type II cells have the ability to rapidly proliferate and differentiate following lung injury.[225, 351] In addition, type II cells can increase in number following exposure to LPS.[352] The proliferative ability and stem-like nature of type II cells makes them a key target for investigating the effects of LPS mediated inflammation in connection with lung carcinogenesis.

Lung cancer is characterized by both genetic and epigenetic changes, or so called "epimutations".[315] Changes in the patterns of DNA methylation can lead to aberrant gene expression, which in turn play a role in carcinogenesis by upregulating protooncogenes and

downregulating tumor suppressor genes.[11] DNA methylation is catalyzed by DNA methyltransferases (DNMT), which methylate the C5 position of cytosine to form 5-methylcytosine (MeC).[11] Methylation of gene promoter sequences typically leads to reduced levels of gene expression.[319] Methylation marks can be removed via active or passive DNA demethylation.[353] Ten Eleven Translocation (TET) family of proteins iteratively oxidize MeC to form 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine.[13-15] Oxidized forms of MeC are also hypothesized to have their own epigenetic functions.[203]

In our previous work, we investigated epigenetic alterations in mouse lung following inhalation exposure to cigarette smoke and LPS (Chapter 4 of this Thesis). We found that methylation and hydroxymethylation status of many genes was affected, including many genes involved in carcinogenesis (**Figure 4.12** in Chapter 4 of this Thesis). However, the earlier study employed whole lung tissue, which is composed of many cell types including Type I and Type II alveolar epithelial cells, lung capillary endothelial cells, and fibroblasts, each with distinct epigenomes. Exposure to cigarette smoke and inflammatory agents may disproportionally affect different cell types. Furthermore, inflammation is associated with significant changes in cellular composition of the lungs, specifically macrophage infiltration and neutrophil accumulation.[354] Therefore, studies utilizing whole tissue are unable to detect cell type-specific epigenetic changes. In the present study, we aimed to investigate the epigenetic responses of alveolar type II cells in the lung towards LPS induced inflammation in the A/J mouse model.

5.2 Materials and Methods

Chemicals and Enzymes

PDE I, PDE II, and DNase I were purchased from Worthington Biochemical Corp. (Lakewood, NJ), while calf intestinal alkaline phosphatase was from Sigma-Aldrich, (Madison, WI). Nanosep10K filters were acquired from Pall corp. (Port Washington, NY). O-(Biotinylcarbazoylmethyl) hydroxylamine was obtained from Cayman Chemical (Ann Arbor, MI). Isotopically labeled D_3-hmC was purchased from Cambridge Isotope Labs (Cambridge, MA). $^{13}C_{10}$,$^{15}N_2$-5-methyl-2′-deoxycytidine and $^{13}C_{10}{}^{15}N_2$-5-formyl-2′-deoxycytidine were synthesized as previously described in Chapters II and IV.[20] All other chemicals used were bought from Sigma-Aldrich (Milwaukee, WI) or Fisher Scientific (Fairlawn, NJ).

Animal Treatments

Female A/J mice were obtained from the Jackson Laboratory (Bar Harbor, ME) and housed in specific-pathogen-free animal quarters at Research Animal Resources, University of Minnesota Academic Health Center. All animal experiments were performed according to the U.S. National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee, University of Minnesota.

***LPS treatment*:** Female A/J mice (8 weeks of age) were divided into 2 groups (7 per group) and treated intranasally under isoflurane anesthesia. Mice in Group 1 were treated with physiological saline (50 µL) every day for 4 days and mice in group 2 were treated with lipopolysaccharide (LPS), 2 µg in 50 µL of physiological saline on day 1 and 5 µg in 50 µL saline on days 2-4. Mice were euthanized in a CO_2 chamber after 4 days of treatment.

Isolation of Type II alveolar cells from lung tissue:

Type II alveolar cells were isolated from whole lung tissues according to published procedures.[352, 355] Lung tissues were perfused with 30 mL of PBS. The whole lung tissue was then injected with 3 mL dispase (5 U/mL, Stem Cell Technology, Vancouver Canada) and placed on ice. Tissues were finely minced, and additional 60 µL dispase (5 U/mL) and 7.5 µL DNase I (2 U/µL, Thermofisher Scientific, Fairlawn, NJ) was added prior to incubation at 37 °C for 45 min, with inverting every 5 min. Dissociated cells were pushed through a 100 micron cell strainer, followed by a 40 micron cell strainer (Corning, Corning, NY) to remove connective tissue. To the resulting cells, RBC lysis buffer (e-Biosciences, San Diego, CA) was added and mixed every 5 min for 30 min at room temperature. The suspension was centrifuged, the supernatant was removed, and the cells were resuspended in 50:50 DMEM/F-12 media. The cells were plated on dual antibody coated plates prepared 24 h in advance with anti-mouse CD16/CD32 and anti-mouse CD45.1 (e-Biosciences). After incubation for 2 hours at 37 °C to facilitate negative selection for type II alveolar cells, the media containing the suspended type II alveolar cells was removed and pelleted

by centrifugation. Cells were lysed and stored at -80 ℃ in 750 µL of Trizol Reagent (Thermo Fisher, Fairlawn, NJ) until DNA/RNA isolation.

HPLC-ESI$^+$-MS/MS quantitation of MeC, hmC, and fC:

Genomic DNA was extracted from type II cells using Trizol extraction protocol according to the manufacturer's protocol. DNA purity and concentration were determined by UV spectroscopy. Genomic DNA (1-2 µg) was subjected to hydrolysis with PDE I (3.6 U, 5 µg), PDE II (3.2 U), DNase I (50U), and alkaline phosphatase (10 U) in 10 mM Tris HCl/15 mM $MgCl_2$ buffer (pH 7) at 37 °C overnight. The hydrolysates were spiked with $^{13}C_{10}{}^{15}N_2$-5-methyl-2′-deoxycytidine (1 pmol), 5-hydroxymethyl-d_2-2′-deoxycytidine-6-d_1 (900 fmol), and $^{13}C_{10}{}^{15}N_2$-5-formyl-2′-deoxycytidine (500 fmol) (internal standards for mass spectrometry synthesized in our laboratory, see Chapters II and IV)[20] and filtered through Nanosep 10K Omega filters (Pall Corporation, Port Washington, NY).

DNA hydrolysates were dissolved in 100 mM ammonium acetate buffer (pH 4.5) containing 100 mM aniline and 400 µM of O-(biotinylcarbazoylmethyl) hydroxylamine (Cayman Chemical. Ann Arbor, MI) and allowed to react for 24 h to derivatize fC to biotinyl-fC.[14] The resulting mixture was dried, followed by offline HPLC to enrich for MeC, hmC, biotinyl-fC, and caC using an Atlantis T3 column (Waters, 4.6 x 150 mm, 3 µm) was eluted at a flow rate of 0.9 mL/min with a gradient of 5 mM ammonium formate buffer, pH 4.0 (A) and methanol (B). Solvent composition was changed linearly from 3 to 30% B over 15 min, increased to 80% over the next 3 min, maintained at 80% B for the

next 2 min, and brought back to 3 % B. The column was equilibrated for 7 min prior to the next injection. dC was quantified by HPLC-UV using calibration curves obtained by analyzing authentic dC standards. HPLC fractions corresponding to MeC, hmC, caC, and biotinyl-fC (7-8.6 min for both hmC and caC, 9-10.5 min for MeC, and 18.7-20.2 min for biotinyl-fC) were combined, dried, and analyzed by isotope dilution HPLC-ESI-MS/MS.

Quantitation of MeC, hmC, and biotinyl-fC was performed using a Dionex Ultimate 3000UHPLC (Thermo Fisher, Waltham MA) interfaced with a Thermo TSQ Vantage mass spectrometer (Thermo Fisher). Chromatographic separation was achieved on a Zorbax SB-C18 column (0.5 x 150 mm, 3 μm, Agilent) eluted at a flow rate of 15 μL/min with a gradient of 2 mM ammonium formate (A) and methanol (B). Solvent composition was maintained at 5% B for the first 3 min and linearly changed from 5 to 40% B for 7 min. Solvent composition was returned to initial conditions (5% B), and the column was re-equilibrated for 4 min. Under these conditions, MeC and $^{13}C_{10}{}^{15}N_2$-MeC eluted at 4.1 min, both hmC and the internal standard d_3-hmC eluted at 3.5 min, while biotinyl-fC and its internal standard ($^{13}C_{10}{}^{15}N_2$-biotinyl-fC) eluted at 7.0 min. Quantitation was achieved by monitoring the transitions *m/z* 258.1 [M + H^+] → *m/z* 141.1 [M – deoxyribose + H^+] for hmC, *m/z* 261.2 [M + H^+] → *m/z* 145.1 [M – deoxyribose + H^+] for D_3-hmC, *m/z* 242.1 [M + H^+] → *m/z* 126.1 [M + H^+] for MeC, *m/z* 254.2 [M + H^+] → *m/z* 133.1 [M + H^+] for $^{13}C_{10}{}^{15}N_2$-MeC, *m/z* 569.1 [M + H^+] → *m/z* 453.3 [M – deoxyribose + H^+] for Biotinyl-5fC, *m/z* 581.2 [M + H^+] → *m/z* 460.4 [M – deoxyribose + H^+] for Biotinyl-$^{13}C_{10}{}^{15}N_2$-5fC. Mass spectrometer were determined by infusion of authentic

standards. Typical settings on the mass spectrometer were: a spray voltage of 2700 V, a sheath gas of 15 units, the declustering voltage was 5 V, and the ion transfer tube was maintained at 350 ℃. The full-width at half-maximum (FWHM) was maintained at 0.7 for both Q1 and Q3. Fragmentation was induced using a collision gas of 1.0 mTorr and a collision energy of 10.3 V.

Histone acetylation analysis:

Type II alveolar cells isolated from the lungs of control and treated mice were homogenized with a glass douncer on ice. Total histones were extracted from the lysates as previously described.[289-290] Histone proteins were acetylated with ($^{13}C_2$,d_3)-acetyl N-hydroxysuccinimide ester to block all unmodified lysines and then digested by trypsin.[290] The peptides were desalted with C18 Stage Tips (3M Corporation, St. Paul, MN) prior to analysis.[356]

Tryptic peptides were analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry (nanoLC-ESI-MS/MS) using a Thermo Scientific Orbitrap Fusion mass spectrometer (Thermo Scientific, San Jose, CA) coupled to a Proxeon Easy nLC 1000 UPLC system (Thermo Fisher Scientific, Odense, Denmark). Each sample was resuspended in HPLC buffer A (0.1% formic acid in water) and loaded onto an in-house packed C18 column (25 cm x 75 μm I.D.) packed with ReproSil-Pur Basic C18 beads (2.5 μm, Dr. Maisch GmbH). Peptides were eluted with a gradient of 5% to 15% B (0.1% formic

acid in acetonitrile) over 26 minutes, then 15% to 35% HPLC buffer B over 16 minutes at 300 nL/min.

Tryptic peptides were analyzed using a FT survey scan from 300-1600 *m/z* at a resolution of FWHM 120,000 (at 200 *m/z*), followed by HCD MS/MS scans using the top speed mode (3 seconds per cycle) at a resolution of FWHM 15,000 (at 200 *m/z*) and a normalized collision energy at 35%. Targeted MS/MS data acquisition was achieved with an inclusion list for fully labeled histone tryptic peptides that covered known lysine acetylation sites. For each peptide, modification isomers with all possible combinations of light/heavy lysine acetylation (Δ M of 42.010565 and 47.036094 Da, respectively) at detectable charge states were considered in the inclusion list for targeted fragmentations.

MS data was searched against the Uniprot Mus musculus proteome database (http://www.uniprot.org) using MaxQuant search engine (v1.4.1.2) as previously described.[357-358] Heavy and light acetylation on lysine as well as methionine oxidation were included as variable modifications, with 6 ppm specified as the precursor mass error and 0.025 Da as the fragment mass error. All peptide spectra matches were filtered at 1% False Discovery Rate with a minimum Andromeda score cutoff of 40. HPLC elution profiles were manually evaluated to ensure accurate quantifications. Only peptides that were confidently identified were selected for stoichiometry analysis. Acetylation stoichiometries of specific sites were calculated using in-house developed scripts based on the extracted peak areas of each modification isomer and quantification of modification-specific fragment ions.[290] Statistical significance analysis of site-specific

acetylation stoichiometry dynamics between control and treated samples was conducted with two-sided Student's t-test using SAS statistical software 9.3 (SAS Institute Inc., Cary, NC).

RNA-Seq of type II cell mRNA:

Type II cells were stored in Trizol (Thermo Fisher) at -80 °Cto minimize RNA degradation. RNA was isolated using Qiagen miRNeasy kits (Qiagen, Hilden Germany) according to the manufacturer's instructions. RNA was quantified using a fluorometric RiboGreen assay (Thermo Fisher). RNA integrity was confirmed using Agilent Bioanalyzer (Agilent, Santa Clara CA). Samples with RNA integrity score higher than 7 were selected for RNA-Seq. Total RNA samples were converted to Illumina sequencing libraries using SMARTer Stranded Total RNA-Seq Kit – Pico Mammalian Input (Takara Bio USA, Mountain View CA). RNA was reverse transcribed into cDNA, and Illumina adapters were added using PCR. rRNA was cleaved with ZapR and R-Probes, and the uncleaved fragments were enriched by PCR. Final library size distribution was validated using capillary electrophoresis and quantified using fluorimetry (PicoGreen) and via Q-PCR. Indexed libraries were normalized, pooled, and size selected to 320bp +/- 5% using Caliper's XT instrument (PerkinElmer, Waltham MA). Truseq libraries were hybridized to a single read flow cell, and individual fragments were clonally amplified by bridge amplification on the Illumina cBot using the HiSeq SR Cluster Kit v4 cBOT (Illumina, San Diego CA). Sequencing was performed on the Illumina HiSeq 2000 sequencing system

using Illumina's SBS chemistry. Primary analysis and de-multiplexing are performed using Illumina's CASAVA software 1.8.2.

Data was imported into CLC Bio Genomics Workbench 7 (CLC Bio, Qiagen, Boston, MA) for quality, mapping, and expression analysis. Principal component analysis was performed using Emperor.[359] Significantly dysregulated genes were clustered with Cluster 3.0, visualized with Java TreeView, and input into Ingenuity Pathway Analysis (Qiagen, Boston MA).

5.3 Results

Epigenetic changes in type II lung cells following intranasal treatment with LPS

Since our earlier data revealed that exposure to LPS induced epigenetic changes in the whole lung (**Figure 4.6** in Chapter IV of this thesis), we hypothesized that epigenetic deregulation is also induced in type II alveolar cells, which possess a high proliferative potential.[225] Alveolar cells Type II cells are considered the cells of origin for non-small cell lung cancer (NSCLC) development.[225] Cell panning methodology was used to isolate type II alveolar cells from lung tissues of mice treated with LPS for 4 days and the corresponding controls. Total RNA and DNA of type II alveolar cells was extracted using Trizol (Thermo Fisher). Global cytosine methylation (MeC) levels in type II cells isolated from control and LPS treated mice was 5.19 ± 0.99 and 5.32 ± 0.71 % of Cs, respectively (**Figure 5.1**). Cytosine hydroxymethylation in genomes of type II cells was decreased in LPS treated mice as compared to controls (0.36 ± 0.08% of Cs vs 0.51 ± 0.26, p = 0.13). Similarly, the levels of fC in type II cells showed a statistically significant decrease in LPS treated animals relative to controls (0.0030 ± 0.002 vs 0.0009 ± 0.0005 % of Cs, respectively, p = 0.039). These data suggest that LPS induced inflammation influences the global levels of hmC and fC in the type II alveolar cells. Similar trend was observed in type II cell depleted lung tissues, although fC changes were not statistically significant (**Figure 5.2**).

Figure 5.1. Global levels of MeC, hmC and fC, in genomic DNA isolated from type II alveolar lung cells of female A/J mice treated intranasally with LPS (4 µg/ day) for 4 days. Data represent mean values ± SD of at least three animals. Statistics were calculated using the Wilcoxon-Mann-Whitney test, p-values: MeC (0.836), hmC (0.366), and fC (0.035).

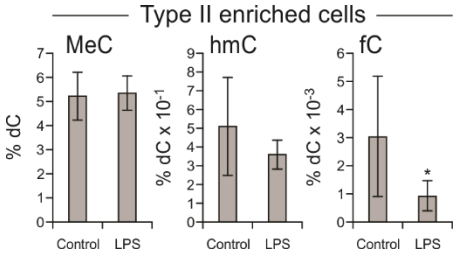
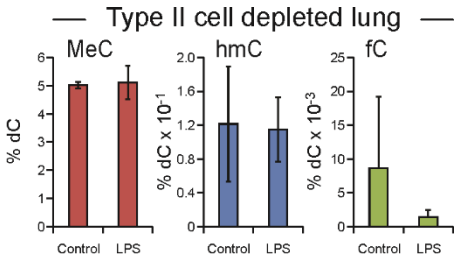


Figure 5.2. Global levels of MeC, hmC and fC, in genomic DNA isolated from type II cell-depleted lung tissues of female A/J mice treated with LPS intranasally (4 µg/ day) for 4 days. Data represent the mean ± SD of at least 4 animals. Statistics were calculated using the Wilcoxon-Mann-Whitney test, p-values: MeC (0.4), hmC (0.4), and fC (0.229).



Gene expression changes in type II cells from LPS treated A/J mice

To determine whether LPS-associated changes in cytosine methylation and hydroxymethylation lead to changes in gene expression, we conducted RNA-Seq analyses using mRNA isolated from the same sample. Total sequencing reads for each sample were within at least 95% of the targeted 20 million reads for each sample. One treated sample had 68.6 million total number of reads, with the rest ranging from 19.3 – 29.1 million reads. Using gene expression profiles, unsupervised principal component analysis was conducted. PCA analysis indicated that control and LPS treated samples separate into two distinct clusters (**Figure 5.3C**) indicative of significant gene expression changes associated with inflammation.

RNA-Seq data was interrogated using the Cuffdiff program available from the Broad Institute.[360] A total of 1,064 genes exhibited altered levels of gene expression (fold change > 2) in Type II cells of LPS treated animals. A subset of dysregulated genes is shown in **Figure 5.3A** as an expression heatmap with clustering. Data for a subset of genes associated with tumor suppression, epigenetic regulation, and the inflammation response is shown in more detail in **Figure 5.3B**. Differentially expressed genes were annotated and analyzed using Ingenuity Pathway Analysis (IPA) to identify key biological processes, pathways (**Figure 5.4A**), and networks affected by LPS treatment. Top diseases and biological functions associated with LPS treatment in type II cells included immunological disease, metabolic disease, immune cell trafficking, cellular movement, and cell to cell signaling and interaction. The most upregulated genes were STFA2, STFA3, SAA3, and

LY6I (**Figure 5.4B**), while the most downregulated genes were CDH4, CHIL4, NXF7, and CYP2A5. The STFA2 and STFA3 genes are both protease inhibitors of papain-like cysteine proteases in the cytoplasm, which prevents the degradation of proteins.[361] SAA3 has been shown to be a protumorigenic mediator in fibroblasts associated with pancreatic cancer[362] and the LY6I protein is associated with inflammation, but its function remains unknown.[363] The downregulated gene CDH4 functions as a tumor suppressor gene in lung cancer[364] and has been shown to be hypermethylated in other cancers.[365] The CHIL4 gene is associated with tissue remodeling and wound healing.[366] NXF7 is part of a family of nuclear export factors and function in mRNA processing in the cytoplasm.[367] CYP2A5 is responsible for bioactivation of carcinogens such as NNK.[368]

Of the 1064 genes dysregulated by exposure to LPS, 859 were identified as being related to cancer, including genes involved in cell proliferation and cell growth. Some examples included matrix metalloproteinase-9 (MMP9), interferon-γ (IFN-γ), and nitric oxide synthase 2 (NOS2).[369] MMP9 assists in breaking down the extracellular matrix to allow for cell growth.[370] NOS2 is an upstream regulator of MMP9 expression, and in increase in NOS2 expression results in an increase in MMP9 expression.[369] IFN-γ is an inflammation response gene responsible for signaling and controlling the immune response.[371] Another large group (N = 269) included genes associated with cell differentiation, such as NOS2 and DNMT3L and proliferation with 405 genes including DAPK1, CDH13, IL21, and MMP9. DNMT3L protein is a key scaffolding protein facilitates the formation of productive complexes with DNMT3A and DNMT3B, which

perform *de novo* DNA methylation.[372] DAPK1 and CDH13 are tumor suppressor proteins that are involved in cell survival and adhesion respectively, and are typically inactivated in lung cancer.[340, 373] IL21 is a cytokine that is induced in the inflammation response and signals $CD8^+$-T cells to expand at the site of inflammation.[374]

Figure 5.3. Effects of LPS induced inflammation in the lungs of A/J mice on gene expression in type II alveolar cells. **A**, Gene expression heat map represented as row z-score and unsupervised hierarchal clustering. **B**, Heat maps showing gene expression changes in tumor suppressor genes, epigenetic regulators, and inflammation-associated genes. **C**, Principle component analysis of the gene expression changes observed in type II cell (red and blue) and in lung tissue of A/J mice following chronic exposure to NNK, LPS, and NNK/LPS.[142]

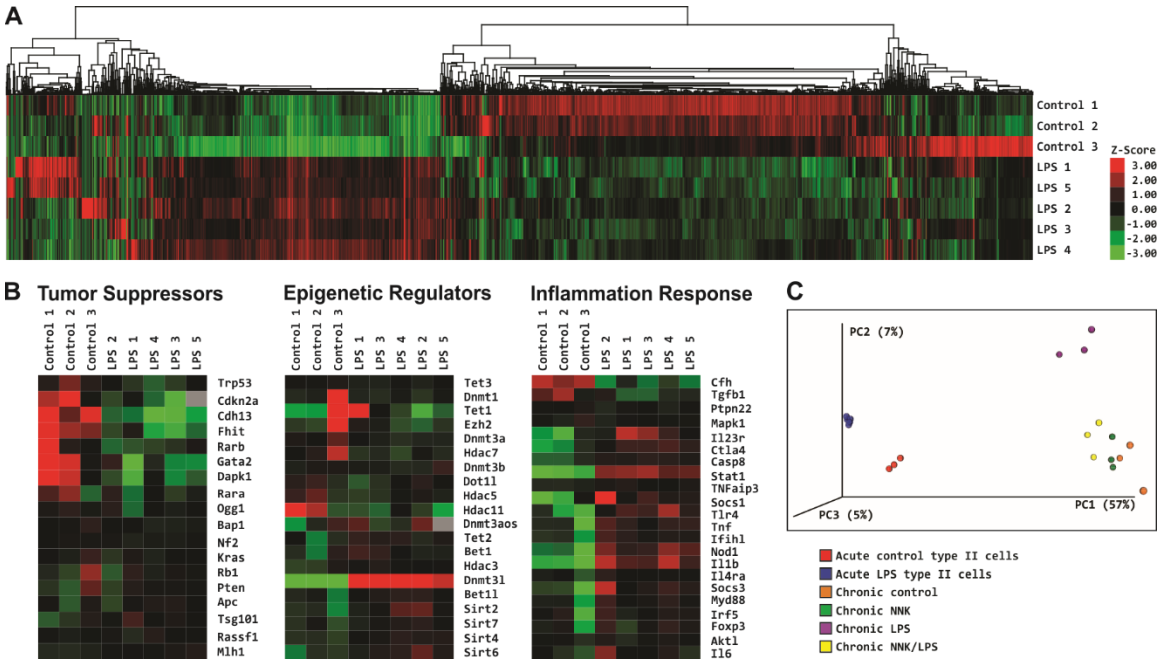
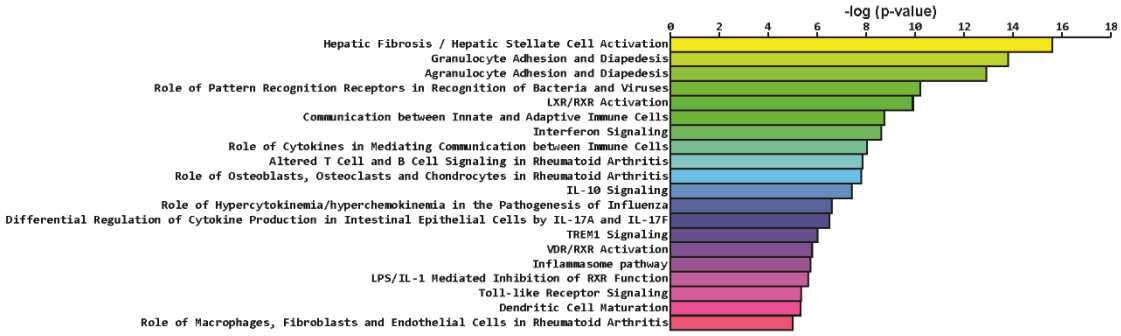


Figure 5.4. LPS-induced gene expression changes in type II alveolar cells of female A/J mice. **A**, Enrichment of canonical pathways of the 1064 differentially expressed genes assessed by IPA software. Top 20 pathways are shown. **B**, Genes showing the largest changes in gene expression.

A



B

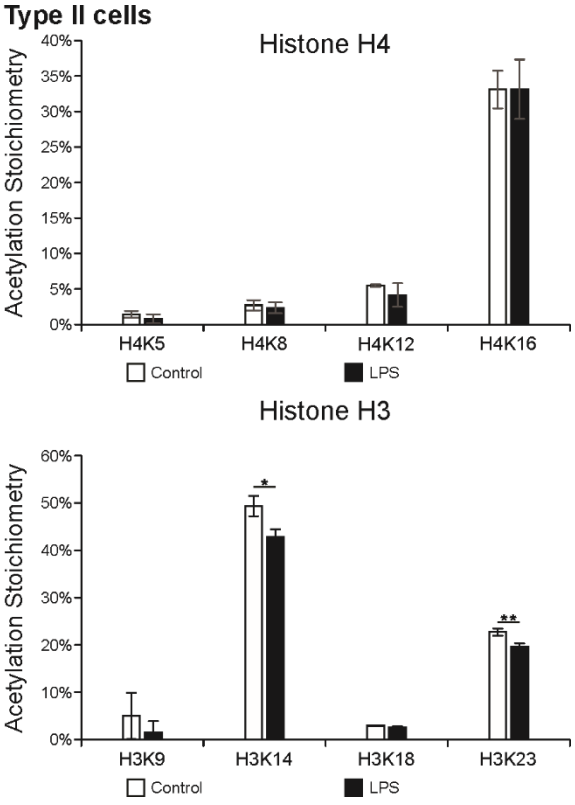
Increased Expression			Decreased Expression		
Protein	**Fold Change**	**p-value**	**Protein**	**Fold Change**	**p-value**
Stfa3	524.14	1.52E-82	Cyp2a5	-20.40	1.23E-11
Gm15056	444.25	3.91E-51	Esr2	-21.00	0.028924
Stfa2	354.61	1.62E-37	Grin2b	-21.36	0.047951
Saa3	344.77	5.27E-19	Cpne5	-21.78	9.92E-11
Ly6i	249.96	3.41E-43	Cd207	-21.87	1.56E-13
BC100530	232.73	3.77E-84	Pcdhac2	-22.28	0.009135
BC117090	207.32	9.05E-50	Cdh4	-28.78	0.032902
2010005H15Rik	159.70	2.4E-12	Fabp1	-31.47	7.18E-14
Gm5416	139.40	5.53E-63	Nxf7	-86.82	0.017719
Gm5483	112.38	5.01E-62	Chil4	-287.30	2.08E-07

Histone acetylation changes in type II cells

Acetylation of histone proteins is a key regulatory mechanism that maintains epigenetic memory in eukaryotic cells.[297] Acetylation of histones H3 and H4 correlates site-specifically with the global gene expression profiles and transcriptional activation,[298-299] while the dysregulation of histone acetylation contributes to tumorigenesis and disease progression in lung cancer.[298, 300] To determine whether LPS treatments induce global changes in histone acetylation in type II cells, we employed a chemical proteomics strategy that enables site-specific identification of the lysine acetylation.[290]

Histone proteins were extracted from type II cells from A/J mice treated with LPS for 4 days and control mice, labeled, and analyzed by nano HPLC-ESI$^+$-MS/MS as described previously.[289] The data were analyzed by in-house developed scripts to calculate site-specific acetylation stoichiometries.[290] Our data indicates that LPS treatment induced global deacetylation of H3K14 and H3K23 in type II cells (**Figure 5.5**), while H3K9, H3K18, and H4 showed no change. Acetylation of H3K14 and H3K23 are strongly associated with active gene expression.[299, 301-304]

Figure 5.5. LPS-induced changes in histone lysine acetylation stoichiometry observed in isolated type II cells of A/J mice intranasally treated with LPS. Data were expressed as percent acetylated and represents mean values ± SD of at least 3 animals. Statistical significance was evaluated between treated and control samples using a two-sided Student's t-test (* $p < 0.05$, ** $p < 0.01$). Histone acetylation stoichiometries were assessed to determine global gene activation by deacetylation due to exposure. Measured acetylation sites include lysines 5, 8, 12, and 16 on histone H4 and lysines 9, 14, 18, and 23 on Histone H3.



5.4 Discussion

The link between inflammation and the development of cancer is well established. Specifically, people with chronic inflammatory bowel diseases, such as ulcerative colitis, Lynch syndrome, and Crohn's disease have an increased risk of colon cancer,[132] and COPD patients are more likely to develop lung cancer.[131, 375] Anti-inflammatory drugs as aspirin may reduce colorectal cancer risk.[313, 376] In 2016, U.S. Preventive Services Task Force recommended the use of aspirin to prevent colorectal cancer among many U.S adults. However, the underlying mechanisms are not completely understood.[131, 377]

In our earlier study, female A/J mice treated with inflammatory agent lipopolysaccharide, exhibited profound epigenetic changes in the lung including altered DNA hydroxymethylation, histone acetylation, and gene expression changes (Chapter IV of this Thesis). This led us to hypothesize that epigenetic deregulation could provide a link between inflammation and lung cancer. In the present study, we aimed to characterize LPS-associated epigenetic changes in Type II alveolar cells, which are thought to be the cells of origin for non-small cell lung cancer due to their high proliferative potential and stem-cell like nature.[225]

Female A/J mice were intranasally treated with LPS for 4 days, and their lung tissues were harvested and subjected to cell panning to enrich for Type II cells. Isotope dilution cap HPLC-ESI-MS/MS was employed to accurately quantify global MeC, hmC, fC in DNA, while histone acetylation was accessed by mass spectrometry based proteomics, and overall changes in gene expression were characterized by RNA seq. While

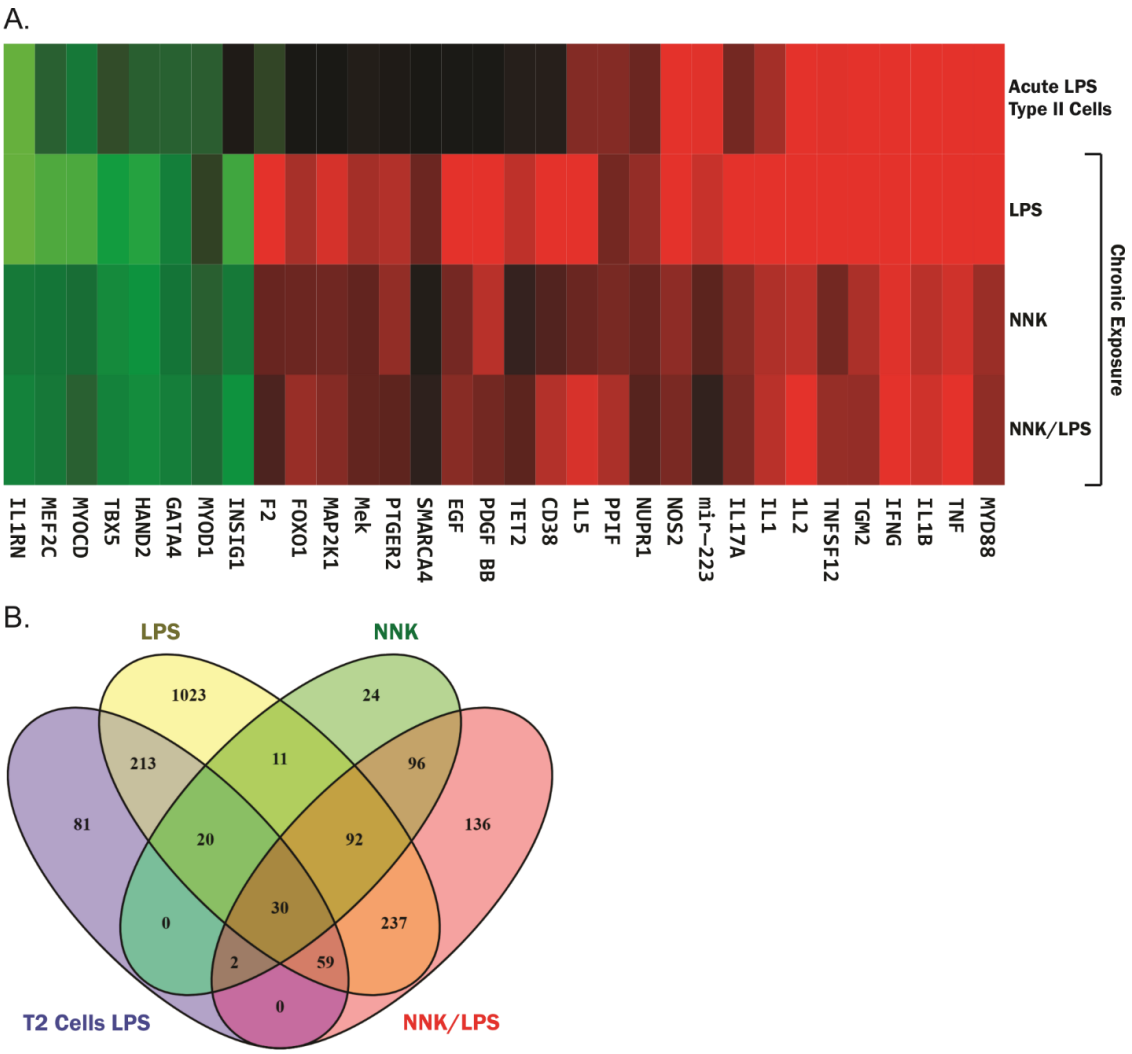
LPS treatment did not affect global levels of MeC in type II cells, a 30% decrease in hmC and 70% decrease in fC was observed (**Figure 5.1**). The type II depleted lung cells showed no change in hmC and a decrease in fC, which was not statistically significant. These results suggest that type II cells are more prone to inflammation mediated epigenetic deregulation the other cell types in the mouse lungs.

RNA-Seq has revealed significant LPS induced gene expression for several tumor suppressor genes associated with lung tumorigenesis, e.g. decreased expression of DAPK1 and CDH13 and increased expression of NOS2 (**Figure 5.3A**). The expression of cytochrome P450 1A1, 2A5, and 2F2 responsible for metabolism of cigarette smoke constituents were increased was decreased. Additionally, expression levels of many proinflammatory genes such as IFN-γ, CCL4, IL21, CXCL9, and CXCL10 were increased (**Figure 5.3B**). Interestingly, DNMT3L, which is a scaffolding protein for DNA methyltransferases DNMT3A/B, showed a significant 18-fold increase in expression upon LPS treatment (**Figure 5.4B**). Overall, the list of deregulated genes was enriched in tumor suppressors, epigenetic regulators, and genes associated with inflammation response. which are consistent with LPS causing epigenetic changes that contribute to the etiology of lung cancer.

Our RNA-Seq data for type II cells was compared with published whole lung data from Qian et al. in the same mouse model treated with LPS and NNK.[142] Principal component analysis (**Figure 5.3C**) of the data sets indicates that the type II cells (red and blue) are distinct from their whole tissue counterparts (orange and purple) and that

exposure to LPS causes a shift up and to the left for both type II cells (blue) and whole tissue (purple) relative to the respective controls (red and orange). To compare the results from isolated cells and whole tissue RNA-Seq, we analyzed the published data alongside the type II cell data using Ingenuity Pathway Analysis. We directly compared upstream regulators for LPS treated type II cells and whole lungs treated with NNK, LPS, or NNK and LPS and found that many of the top regulators are indeed the same, such as MYD88, NOS2, TNF and many interleukins (**Figure 5.6A**). We also noted that the isolated type II cells shared many upregulated genes with lung tissue exposed to chronic LPS (**Figure 5.6B**). Even though there were many similarities, there were also many upstream regulators that were identified in whole tissue experiments that were not considered upstream regulators in type II cells by IPA including TET2, EGF, MEK, and FOX01 (**Figure 5.6A**). Taken together, the type II cells isolated from mice treated with LPS for a single week present many characteristics similar to long term inflammation exposure mediated by LPS and these gene expression changes occur prior to large changes in epigenetics at the global level. These changes in gene expression of type II cells indicate that the effects of inflammation on transcription occur rapidly in these important cells and that inflammation could contribute to an altered epigenetic state pushing the type II cells toward a cancerous phenotype.

Figure 5.6. Comparison of RNA-Seq data obtained in this study (type II cells) to previously published data of Qian et al. exploring chronic exposure to NNK, LPS, or NNK and LPS in whole lung tissue.[142] (**A**) Ingenuity Pathway Analysis of upstream regulators increased or decreased in type II cells and whole lung chronic treatment. (**B**) Venn diagram of the significantly upregulated genes (4-fold increased and $q < 0.05$)) in the combined data set with type II cells with LPS (T2 cells LPS) and whole lung RNA-Seq after chronic administration of LPS, NNK, or NNK and LPS.



To understand how gene expression changes in the lung could be influenced by changes in site-specific DNA methylation and hydroxymethylation, we compared RNA-Seq expression data to RRBS data described in Chapter IV of this Thesis. RRBS and oxoRRBS were used to determine the site-specific methylation and hydroxymethylation levels in lung DNA from A/J mice exposed to environmental tobacco smoke for 10 weeks or controls. These data are based on sequencing of DNA from whole lung tissue and represent a heterogenous lung cell population compared to isolated type II cells used for RNA-Seq analysis. In addition, the DNA used for RRBS data collection come from mice exposed to cigarette smoke, a complex mixture of carcinogens and co-carcinogens, while RNA-Seq was performed on cells from mice treated with LPS. The differences between the sample composition and treatment strategy allow the drawing of tenuous conclusions.

To compare the RRBS and RNA-Seq data, the changes in gene expression observed in type II cells were associated with changes in DNA epigenetic marks, genes that exhibited altered DNA methylation and hydroxymethylation in A/J mice treated with environmental tobacco smoke for 10 weeks[140, 291] (**Figure 4.2** in Chapter II of this thesis) were plotted against the gene expression heat map for type II cells of A/J mice treated with LPS. Using this analysis, genes exhibiting increased methylation show an enrichment for genes with reduced transcription (**Figure 5.7A**). Similarly, genes with increased hmC also show an enrichment for genes with reduced transcription (**Figure 5.8A**). For both set of genes having decreased levels of MeC or hmC, there is little enrichment for gene expression changes showing an increase or decrease (**Figure 5.7B** and **Figure 5.8B**, respectively). A

correlation of DNA methylation and reduced gene expression is well known in the literature.[318, 336-337, 339] More interesting are the increases in hydroxymethylation, which shows an enrichment for reduced gene expression, while the decreases in hmC show little enrichment for changes in expression (**Figure 5.8**). The lack of clear association could be due to the prevalence of hmC in promoters and enhancer sequences, affecting transcription in a different way than MeC.[378]

Figure 5.7. Correlation of type II cell gene expression (**Figure 5.3A**) and genes with altered methylation identified by oxoRRBS in A/J mice exposed to cigarette smoke (**Figure 4.2A** from Chapter IV of this thesis). The top 100 increases and decreases in MeC were correlated with type II cell expression. **A**, Genes with increased levels of MeC (labeled on the right) show an enrichment for reduced gene expression in type II cells. **B**, Genes with decreased MeC (labeled on the right), show an enrichment for decreased gene expression in type II cells.

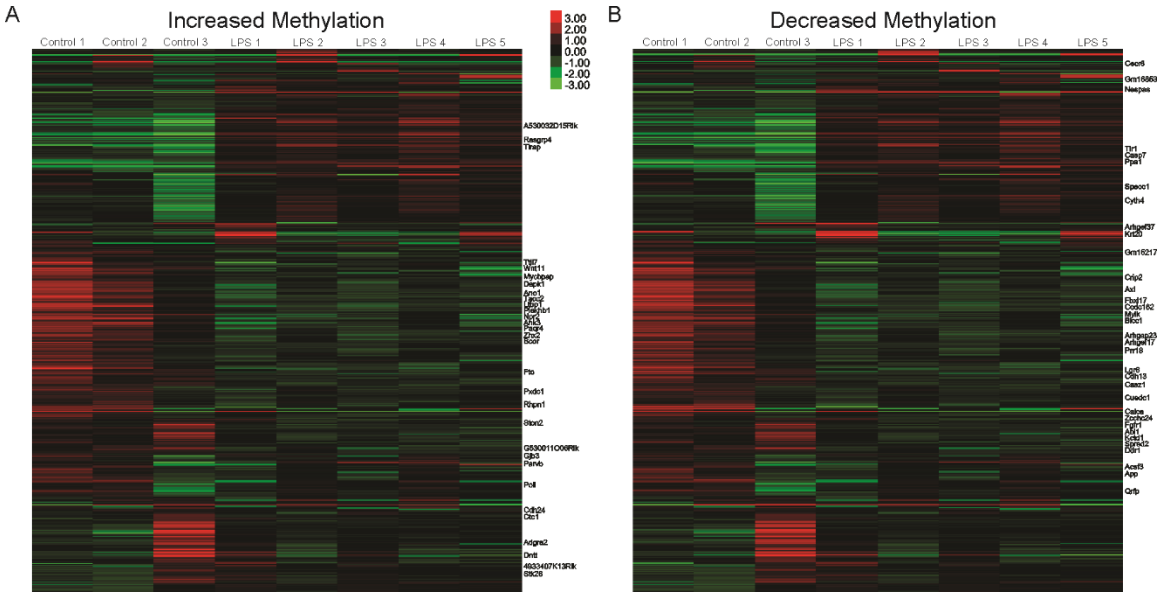
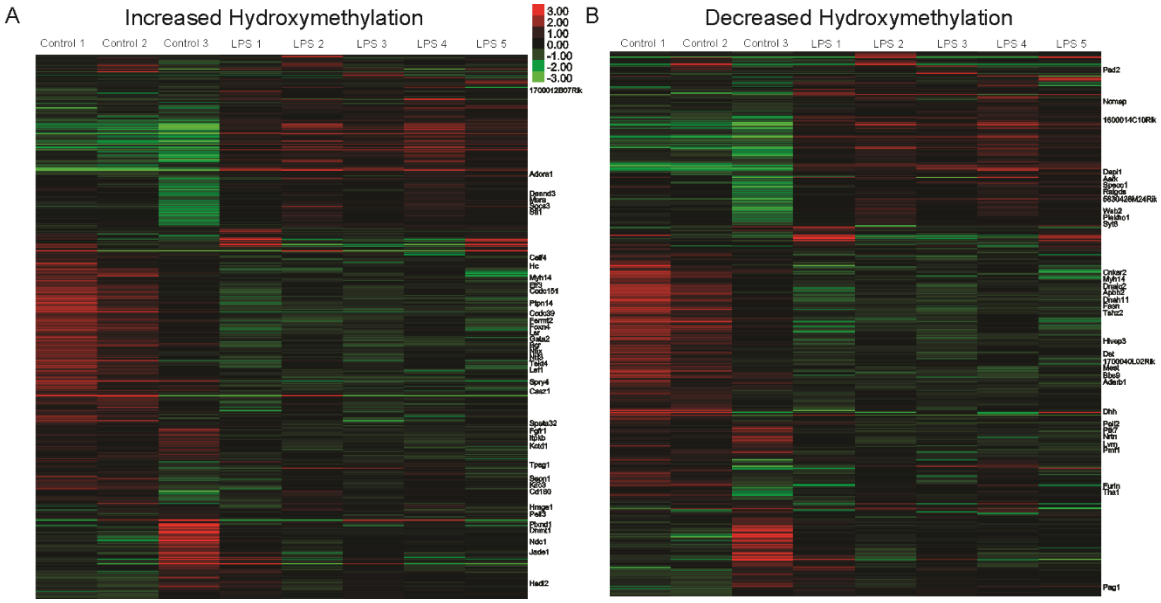


Figure 5.8. Correlation of type II cell gene expression (**Figure 5.3A**) and genes with altered methylation identified by RRBS and oxoRRBS in mice exposed to cigarette smoke (**Figure 4.2A** from Chapter IV of this thesis). The top 100 increases and decreases in hmC were correlated with type II cell expression. **A**, Increased amounts of hmC (labeled on the right) are enriched for genes with reduced gene expression (heatmap). **B**, Decreasing levels of hmC (labeled on the right) show no enrichment for the gene expression changes in type II cells.



In summary, we elucidated the effects of LPS on the global levels of DNA epigenetic marks, histone acetylation, and transcriptome of type II cells from A/J mice. We found that short term exposure to LPS reduces the levels of hmC and fC in type II cells, which occurs in conjunction with the loss of H3K14 and H3K23 acetylation. Our study also explored the transcriptomic changes induced by acute LPS exposure and found tumor suppressor proteins CDH13 and DAPK1 having reduced expression. The most enriched upstream regulators in type II cells are the same as those observed following chronic exposure to LPS (**Figure 5.6**). Taken together these findings support type II cells as the sites of lung carcinogenesis and indicate epigenetic changes as a potential mechanism of type II cell driven lung carcinogenesis.

VI. AFFINITY BASED PROTEOMICS TO IDENTIFY THE READERS OF EPIGENETIC DNA MODIFICATIONS IN HUMAN LUNG

Seiler, C.L.; Fernandez, J.; Rajczewski, A.; Trisko, P.; Andersen, M.; Han, Q.; and Tretyakova, N.Y.

This is a collaborative work between Christopher Seiler, Jenna Fernandez, Andrew Rajczewski, Paul Trisko, Molly Andersen, and Qiyuan Han under the direction of Professor Natalia Tretyakova. Christopher Seiler performed sample preparation and LC-MS/MS and LC-MS/MS/MS of TMT-tagged proteomics with the assistance of Jenna Fernandez, Andrew Rajczewski. Paul Trisko, and Molly Andersen. Christopher Seiler analyzed the MS data using PD 1.4, 2.1, Scaffold, Cluster 3.0, and Java Tree View. Christopher Seiler and Paul Trisko utilized the DAVID informatics platform for analysis of protein readers. Christopher Seiler, Jenna Fernandez, Andrew Rajczewski, Paul Trisko, and Molly Andersen developed the methodology described herein. Christopher Seiler, Paul Trisko, and Molly Andersen synthesized and purified original DNA capture sequence. Qiyuan Han laid the ground work for and provided advice for developing the PCR amplification techniques for producing longer DNAs as protein pull-down baits.

6.1 Introduction

Epigenetic control of gene expression is critical in cellular differentiation and is required for the maintenance of cellular identity. 5-methylcytosine (MeC) is the most abundant and the most widely studied epigenetic modification in DNA.[11] The presence of MeC in promoter regions of genes is generally associated with gene repression, while cytosine methylation in gene bodies can be associated with active transcription.[11] In 2009, it was discovered that MeC can be oxidized to 5-hydroxymethyl-2′-deoxycytidine (hmC) by the Ten Eleven Translocation (TET) family of proteins.[13] hmC can be further oxidized to 5-formyl-2′-deoxycytidine (fC) and 5-carboxyl-2′-deoxycytidine (caC) (**Figure 1.2 in Chapter I of this Thesis**).[15-16] The oxidized forms of MeC (oxiMeC) can be enzymatically removed by Srap1 endonuclease (SOS associated peptidase 1).[21] In addition, fC and caC can both be excised by thymine DNA glycosylase (TDG) and subsequently replaced with C, leading to active demethylation and gene reactivitation.[16]

In addition to their role as demethylation intermediates, oxiMeC have been hypothesized to possess unique epigenetic functions.[14] This hypothesis is supported by recent studies identifying unique protein readers of MeC, hmC, fC, and caC in mouse embryonic stem cells (ESC) and in brain cells using synthetic DNA constructs and PCR amplicons.[26-27] However, cellular proteomes vary greatly across different cell types, and protein readers of MeC, hmC, fC, and caC in tissues other than brain and cell types other than ESC have not previously been investigated. Because of the importance of epigenetic

changes in lung cancer etiology[335-336, 340, 344] proteins interacting with epigenetic DNA marks in the lung are of interest.

In the present report, a mass spectrometry-based approach was used to identify protein readers of MeC, hmC, fC, and caC in normal human bronchial epithelial cells. Synthetic DNA constructs derived from the WTH3 gene and containing MeC or its oxidized variants was used as a DNA "bait" to pull down specific protein readers from nuclear extracts from human bronchial epithelial cells, and the protein binders were identified and quantified using based mass spectrometry-based proteomics with TMT tags.[379]

6.2 Materials and Methods

Materials

All nucleoside phosphoramidites, solvents, and solid supports for solid phase DNA synthesis were purchased from Glen Research Corporation (Sterling, VA). Pre-cast SDS-PAGE gels, Pierce™ BCA Protein Assay kit, Pierce™ Quantitative Colorimetric Peptide Assay, Pierce™ Quantitative Colorimetric Peptide Assay, Pierce™ C18 spin columns, Pierce™ High pH Reversed-Phase Peptide Fractionation Kit, and TMT 6-plex reagents were purchased from Thermo Fisher Scientific (Waltham, MA). Mass spectrometry grade water and acetonitrile were purchased from Fisher Scientific (Hampton, NH). Mass spectrometry grade formic acid, nuclear extraction kit, and Amicon 3K filters were purchased from Millapore-Sigma (Burlington, MA).

Solid phase DNA synthesis and purification

The synthetic DNA sequences used in this study were derived from the WTH3 promoter, sequence 5′-**B**TGTTTTCAC**X**GCACCATTGTTTTTAGTACATATGTT-3′ (where B = biotinylated thymidine, and X = C, MeC, hmC, fC, or caC, **Table 6.1**). Oligodeoxynucleotides containing modified cytosines were synthesized on an ABI 394 automated DNA synthesizer (Applied Biosystems, CA) according to the manufacturer's recommendations. Synthetic DNA was deprotected and cleaved from solid support according the manufacturer's directions. To increase DNA yields, the solid support was sonicated for 30 min twice in water.

Synthetic DNA oligodeoxynucleotides were purified by reverse phase HPLC on an Agilent 1100 system interfaced with UV-VWD detector and equipped with an autosampler and automated fraction collector. DNA strands containing Biotin-T were purified using a Varian Pursuit C18 column (250 mm x 10 mm x 5 μm, Supelco, Bellafonte, PA). The column was eluted at 40 ℃ with a gradient of 100 mM triethylammonium acetate, pH 7 (A) and 100% ACN (B) at a flow rate of 3 mL/min. A linear gradient of 12-22.5% B over 36 min was employed, and the column was equilibrated for 13 min. DNA strands of interest typically eluted between 21-22 minutes.

For DNA strands which did not contain biotin, a Synergi Hydro RP column (250 mm x 10 mm x 4 μm, Phenomenex, Torrance, CA) was used. The DNA were purified using a gradient of 100 mM triethylammonium acetate, pH 7 and ACN. The column was eluted at 3 mL/min with a linear gradient from 10 to 22% B over 30 min, followed by re-equilibration for 15 min. DNA strands of interest typically eluted at 15 min. Following purification, the DNA was desalted using NAP-5-Illustra™ size exclusion columns (GE Healthcare, Piscataway, NJ) according to the manufacturer's protocols. The presence of MeC, hmC, fC, or caC in DNA strands was confirmed by capillary-HPLC-ESI⁻MS, and DNA concentrations were determined from the amount of 2′-deoxyguanosine in enzymatic digests using previously published HPLC-UV protocols.[239-241]

Double stranded DNA was obtained by combining equal amounts of the complementary strands in 10 mM Tris, pH 8.0, and 50 mM NaCl buffer. DNA was heated to 90 ℃ for 5 min and then allowed to slowly cool to room temperature.

Table 6.1. Nucleotide sequences and mass spectrometry characterization of synthetic DNA 37-mers employed for affinity proteomics experiments. Sequences were derived from WTH3 gene promoter using UCSC genome browser.[380] b = biotin. Molecular weights were obtained from HPLC-ESI$^-$-MS/MS on an Agilent MSD ion trap operating in the negative ion mode**.**

Strand	Sequence 5'-3'	Expected Mass (Da)	Experimental Mass (Da)
Biotin-C	bTTGTTTTCACCGCACCATTGTTTTTAGTACATATGTT	11,659.0	11,658.9
Biotin-MeC	bTTGTTTTCAC**MeC**GCACCATTGTTTTTAGTACATATGTT	11,674.2	11,674.0
Biotin-hmC	bTTGTTTTCAC**hmC**GCACCATTGTTTTTAGTACATATGTT	11,688.3	11,688.9
Biotin-fC	bTTGTTTTCAC**fC**GCACCATTGTTTTTAGTACATATGTT	11,686.2	11,686.9
Biotin-caC	bTTGTTTTCAC**caC**GCACCATTGTTTTTAGTACATATGTT	11,702.3	11,702.0
C	AACATATGTACTAAAAACAATGGTGCGGTGAAAACAA	11,449.3	11,449.0
MeC	AACATATGTACTAAAAACAATGGTG**MeC**GGTGAAAACAA	11,464.3	11,464.1
hmC	AACATATGTACTAAAAACAATGGTG**hmC**GGTGAAAACAA	11,480.3	11,480.0
fC	AACATATGTACTAAAAACAATGGTG**fC**GGTGAAAACAA	11,478.2	11,478.0
caC	AACATATGTACTAAAAACAATGGTG**caC**GGTGAAAACAA	11,494.3	11,494.1

Nuclear extract preparation

Human bronchial epithelial cells (HBEC) were cultured in keratinocyte serum free medium to 70-80% confluency at 37 ℃ with 5% CO_2. Cells with a maximum passage number of 9 were harvested using trypsin. The cells were washed with ice cold PBS three times, and nuclear extracts were prepared with a Nuclear Extraction Kit (Millipore-Sigma, Burlington, MA) according to the manufacturer's instructions.

Nuclear extract protein concentrations were determined using Pierce BCA Protein Assay Kit. Extracts were used immediately or aliquoted, snap frozen in dry ice/acetone and stored at -80 ℃ for up to three months.

Protein pull-downs to identify specific readers of oxiMeC

Preparing the streptavidin beads

Streptavidin beads (450 µL, GE Healthcare, Piscataway, NJ) were washed three times with 9 mL of freshly prepared DNA Binding Buffer (DBB, 1M NaCl, 10 mM Tris, 1 mM EDTA, 0.0005% NP40, pH 8.0). Following the third wash, the beads were re-suspended in 10.5 mL of DBB and split into two equal aliquots, the first one for binding biotinylated DNA and the second one for clearing the nuclear extract of biotinylated molecules and proteins.

Removing endogenously biotinylated molecules and proteins from the nuclear extract

The first aliquot was spun down to remove DBB was removed, and the beads were resuspended in 9 mL of Protein Binding Buffer (PBB, 150 mM NaCl, 50 mM Tris, 1 mM DTT, 0.0025% NP40, pH 8.0). To this first aliquot was added 12 mg of nuclear extracted proteins. The beads and protein mixture were inverted at 4 ℃ for 2 h to remove any endogenously biotinylated molecules. Following incubation, the supernatant was removed and placed into a clean Eppendorf tube and stored until use in affinity purification.

Binding DNA to the beads

The second aliquot of beads from above was divided into 5 equal portions. To each aliquot of beads, 19.5 nmol (3 x 6.5 nmol) of the corresponding biotinylated dsDNA (C, MeC, hmC, fC, or caC) was added (**Table 6.1**). DNA was incubated with beads for 1 h at room temperature with continuous inverting. Following incubation, DNA-functionalized beads were divided into 3 equal parts and centrifugated on a table-top centrifuge to remove the supernatant. The beads were resuspended in 300 µL of PBB.

Capture of specific protein readers

An aliquot of HBEC nuclear protein extract (800 µg), which was cleared of endogenously biotinylated molecules as described above, was added to each sample. Proteins were incubated with DNA bound to beads for 2 h at 4 ℃. After the incubation, the beads were washed three times with 300 µL of PBB. Protein binders were eluted by adding 0.1% SDS and heating the sample at 90 ℃ for 10 min. The sample was briefly

centrifuged, and the supernatant was transferred to an Amicon 3K filter for processing. A detailed protocol for the pull-down methodology can be found in **Appendix A4**.

Reduction, alkylation, and protein digestion using filter aided sample preparation (FASP) methodology

To prepare the proteins for mass spectrometry analysis, the eluted proteins, which were loaded onto Amicon 3K filters, were centrifuged at 14,000g for 10 min, washed three times with 100 µL 50 mM HEPES (pH 8), and resuspended in 100 µL 50 mM HEPES. To the solution, 1 µL of 200 mM DTT/0.1% SDS in 50 mM HEPES was added and the incubated for 1 hour at 55 ℃. Immediately before use, 9 mg of iodoacetamide was dissolved in 132 µL of water to a concentration of 375 mM. To each pulled-down protein sample, 1 µL of the iodoacetamide solution was added and the proteins were incubated for 30 min at room temperature in the dark. Following the alkylation, samples were centrifuged at 14,000g for 10 min. Samples were washed three times with 100 µL of 50 mM HEPES. The proteins were suspended in 50 µL of 50 mM HEPES and incubated with 1 µg of trypsin overnight at 37 ℃. The resulting tryptic peptides were recovered by centrifugation at 14,000g for 10 min into a clean Amicon collection tube. Amicon 3K filters were washed three times with 50 µL 0.1% formic acid added and centrifuged at 14,000g for 10 min to ensure peptides did not stick to the filter. An additional centrifugation was performed with 50 µL of 0.1% formic acid in 70% ACN. A protocol for the FASP methodology can be found in the **Appendix A4**.

Gel-based sample preparation (alternative to FASP)

Proteins eluted from streptavidin beads were run into an SDS-PAGE gel for 10 min at 140 V. The top 1 cm of gel, containing protein was diced into 1 mm cubes. The gel pieces were washed with water and then twice with a 1:1 mixture of 25 mM ammonium bicarbonate and ACN. The gel was dehydrated with ACN for 30 s. The gel pieces were rehydrated in 25 mM ammonium bicarbonate containing 10 mM DTT and incubated at 56 ℃ for 1 hour. After incubation, the supernatant was removed and replaced with 55 mM iodoacetamide in 25 mM ammonium bicarbonate. The gel pieces were incubated for 30 min at room temperature in the dark to alkylate the free cysteines. The gel pieces were washed twice with 1:1 25 mM ammonium bicarbonate and ACN, and then dehydrated with ACN for 30 s. The gel pieces were rehydrated with 25 mM ammonium bicarbonate containing 1 µg of Trypsin and digested overnight at 37 ℃. Peptides were extracted from the gel pieces by two incubations with 50% ACN, 0.3 % formic acid, and another incubation with 80% ACN, 0.3% formic acid.

TMT labeling of tryptic peptides

The eluted peptides were dried and reconstituted in 5% ACN/0.5% TFA. The reconstituted peptides were desalted using Pierce™ C18 Spin Columns according to the manufacturer's instructions. After desalting, the peptides were quantified using Pierce™ Quantitative Colorimetric Peptide Assay. Equal peptide amounts (2.5 µg) were aliquoted

to a fresh tube and labeled with TMT-sixplex reagent (Thermo-Fisher) in 30% ACN in 50 µL total volume. The samples were incubated for 2 h at room temperature. The reactions were quenched by the addition of 4 µL of 5% hydroxylamine followed by 15 min incubation. Differentially tagged samples were combined and concentrated to dryness on speed vac. The combined peptide samples were fractionated using Pierce™ High pH Reversed-Phase Peptide Fractionation Kit according to the manufacturer's instructions for TMT-labeled peptides. The resulting 8 fractions were concatenated to 4 fractions and the dried prior to LC-MS/MS. A detailed protocol for the TMT labeling can be found in the **Appendix A4**.

NanoLC-ESI$^+$-MS/MS Identification and Relative Quantitation of Protein Readers

TMT-labeled peptides were reconstituted at a concentration of 0.5 µg/µL in 5% ACN and 0.1% formic acid. Peptides were analyzed with an LTQ Orbitrap Fusion Tribrid mass spectrometer interfaced with a Dionex Ultimate3000 HPLC system. The samples were loaded onto a pulled-tip fused silica column with a 100 µm inner diameter packed in-house with 20 cm of 3 µm ProntoSIL-C18AQ resin (120 Å, Bischoff Chromatography, Leonberg, Germany) that served both as the analytical column and as a nanospray ionization emitter. HPLC flow rate was maintained at 1 µL/min for the first 5.5 minutes, reduced to 300 nL/min for the next 132 min, and returned to 1 µL/min for the final 7.5 min of the run. HPLC solvents were comprised of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Solvent composition was held at 5% solvent

B for 6 min, followed by an increase to 7% over 2 min, 25% over 99 min, 60% over 20 min, and finally to 95% over 1 min. The solvent composition was held at 95% for 8 min, and returned to 5% over 1 min and re-equilibration for 8 min.

The peptides were introduced to the mass spectrometer using a spray voltage of 2200 V, with an ion transfer tube temperature of 350 °C, and an RF Lens value of 60%. The mass spectrometer was operated in top speed mode (3 s), in which the instrument collected as many unique MS^2 spectra as possible during the 3 s time window. The method utilized data dependent acquisition with dynamic exclusion (repeat count: 1, exclusion duration: 15 s). For every scan cycle, one full MS scan (m/z 320–2000) was collected at a resolution of 120,000 with an automatic gain control (AGC) target value of 2×10^5, followed by MS^2 scans of as many dependent scans as possible within a cycle time of 3 s with HCD (normalized collision energy = 38%, isolation width = 0.7 m/z, resolution = 30,000) at an AGC target value of 1×10^5. Ions with a charge state of 2-7 were included from the analysis.

Data analysis

Peptide sequences were determined from their MS^2 spectra using Proteome Discoverer 2.1 (Thermo Fisher Scientific, Waltham, MA) using the Uniprot human protein database.[381] Precursor ion *m/z* mass tolerance was set at ± 10 ppm, with fragment ion *m/z* tolerance at ± 0.06 Da, and up to two missed cleavages were allowed. Static modifications included carbamidomethylation of cysteine (+57.021 Da) and TMT 6-plex (+219.163 Da) groups on lysines and peptide N-termini. Dynamic modifications included oxidation

(+15.995 Da, M), deamidation (+0.984 Da, N and Q), acetylation (+42.011 Da, K and S), and phosphorylation (+79.966 Da, S, T, and Y). Identified peptides were filtered using a 1% false discovery rate for high confidence and a 5% false discovery rate for medium confidence. At least 1 unique peptide fragment was required for identification. Isotope impurities for all six TMT reporter ions, provided by the manufacturer, were included in the analysis.

From the identified peptides, the relative affinity of a protein for a corresponding DNA bait is proportional to the TMT tag abundance in the multiplexed sample. To allow for relative quantitation of pulmonary proteins with affinity for MeC, hmC, fC, and caC in DNA, TMT-sixplex reporter ion intensities were extracted by Proteome Discoverer 2.1. The mass tolerance for the extracted ion was set to 0.01 Da, and the signal to noise ratio of the reporter ion was at least 10. Top 3 peptides were used to establish the relative protein abundances. Quantitation was performed relative to the TMT-tagged reference sample, which served as an internal control and was created by mixing equal amounts of all 15 samples to facilitate comparisons of protein levels across multiple LC-MS/MS runs and replicates.

Bioinformatics analysis

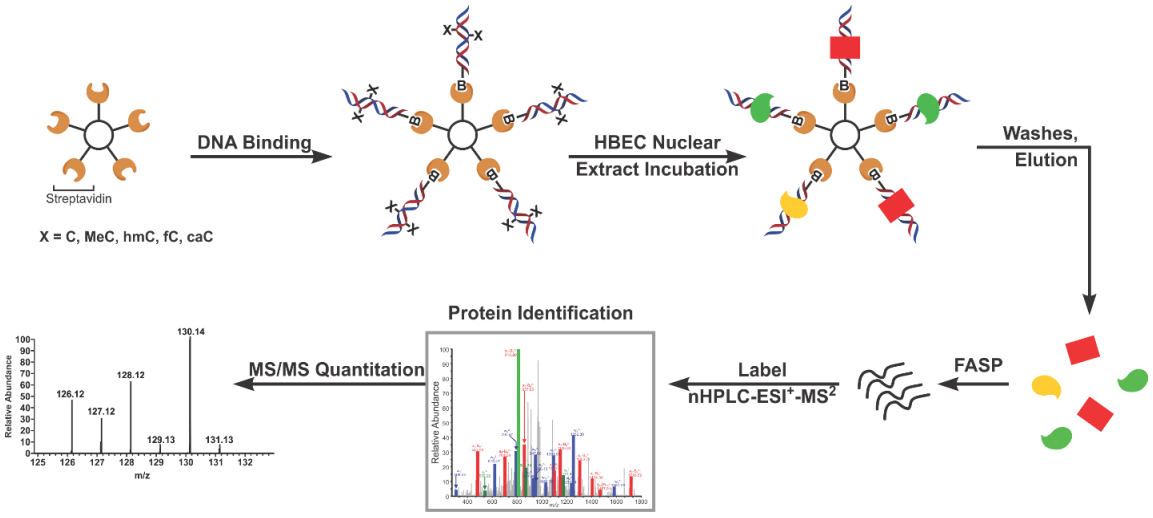
Protein abundance data was processed by Ingenuity Pathway Analysis (IPA, Qiagen, Hilden, Germany), the Database for Annotation, Visualization, and Integrated Discovery (DAVID),[382-383] and Cluster 3.0. Proteins with increased affinity were input into

IPA and a core expression analysis was performed. The core expression analyses were compared across groups using the comparison analysis tool. Whole protein TMT abundance data was clustered using Cluster 3.0 using average linkage. Clustering analysis results were visualized with Java TreeView.

6.3 Developing methodology to profile protein readers of oxiMeC

DNA sequence employed in these experiments was derived from the WTH3 promoter and contained one site-specific cytosine modification at a centrally located CpG site (5′- bTTGTTTTCAC**X**GCACCATTGTTTTTAGTACATATGTT -3′, where b = biotin, and **X** = C, MeC, hmC, fC, or caC, **Table 6.1**). DNA strands were synthesized using solid phase synthesis with commercial nucleoside phosphoramidites. Biotinylated DNA strands were attached to streptavidin coated beads, which were subsequently incubated with nuclear extract from human bronchial cells to capture protein readers of each DNA epigenetic mark (**Scheme 6.1**).

Scheme 6.1. Schematic representation of the affinity pull-down procedure. Streptavidin coated beads are incubated with biotinylated DNA. DNA functionalized beads are incubated with nuclear protein extracts from human bronchial epithelial cells (HBEC). Weak protein binders were washed away, while strongly binding proteins were eluted with SDS/heat. To identify protein readers of epigenetic marks and to quantify their relative affinity for MeC, hmC, fC, and caC, proteins were digested with trypsin and labeled with TMT isotope tags. The resulting peptides were identified and quantified using nano HPLC-NSI-MS/MS on an Orbitrap Fusion mass spectrometer.



Initial tests of our affinity capture methodology were performed using three DNA duplexes containing a central CG, MeCG, or caCG dinucleotide (**Figure 6.1**). In this experiment (Experiment 1, **Table 6.2**), 400 μg of protein was incubated with 4 nmol of DNA bound to 30 μL of streptavidin coated beads. Captured proteins were eluted with 100 mM sodium chloride solution heated at 90 °C for 20 min and run into a SDS-PAGE gel. Proteins were imaged using Coomassie Blue staining (**Figure 6.1**). From the gel image, we observed weak protein bands at the middle range of the gel, darker bands in the high molecular weight region, and a strong band around 13 kDa corresponding to streptavidin (**Figure 6.1**). The proteins were subjected to an in-gel digestion, followed by dithiothreitol treatment to reduce disulfide bonds, and iodoacetamide to alkylate the free cysteines (**Figure 6.1**). The resulting peptides were analyzed by nanoLC-NSI-MS/MS on an Orbitrap Velos mass spectrometer, and the proteins were identified using TINT proteomics software pipeline at the Minnesota Supercomputing Institute. The most abundant proteins detected were keratins, streptavidin, and trypsin, which are all artifacts of sample preparation. In addition to protein artifacts, we were able to identify some lower abundance proteins which were found only in the sample prepared from the caC pull-down (**Table 6.3**). This initial experiment showed that our protocol was capable of capturing DNA binding proteins, and their identification was possible. However, very few meaningful proteins were detected, suggesting that the procedure needed to be optimized. Furthermore, the requirement for separate LC-MS/MS runs required for each pull-down limited our ability to perform relative quantitation of proteins that bind to more than one epigenetic mark.

Figure 6.1. PAGE- of proteins pulled down using the WTH3 DNA duplexes containing C, MeC, and caC. Gel was gut into 5 MW fractions prior to tryptic digestion and LC-MS/MS analysis.

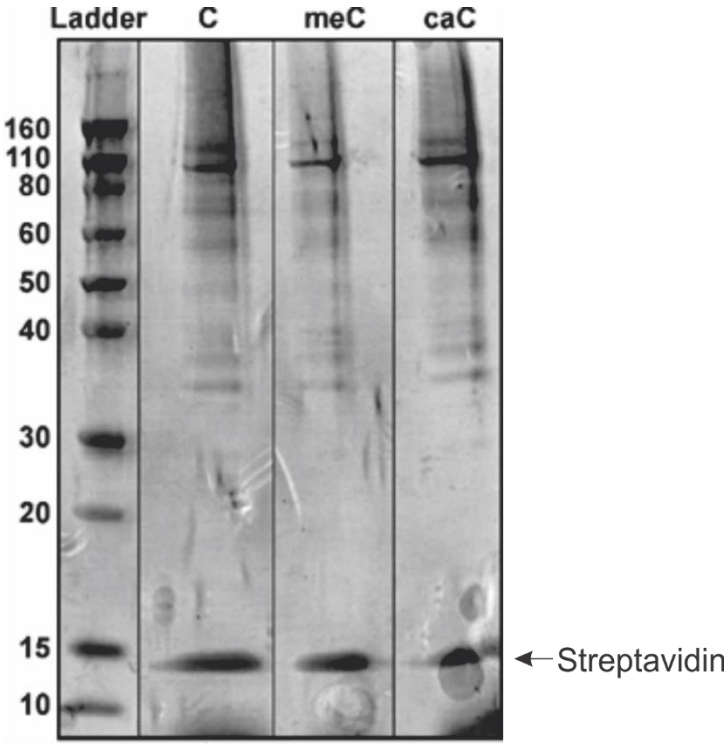


Table 6.2. A summary of affinity pull-down experiments.

	Input		Output	
Experiment	**DNA**	**Protein Amount**	**Peptides (µg)**	**Protein IDs**
1	4 nmol	400 µg	ND*	ND*
2	4 nmol	400 µg	ND*	ND*
3	4 nmol	400 µg	7-9	28
4	4 nmol	800 µg	2.8	371
5	5 nmol	800 µg	2.8	ND*
6	2 nmol	200 µg	4	61
7	3 nmol	900 µg	ND*	288 (MeC), 146 (hmC)
8	3 nmol	900 µg	ND*	74
9	3 nmol	900 µg	ND*	ND*
10	NA**	NA**	2.5	1600
11	50 ng	75 µg	1.7	86

* ND – Amounts were not determined.
** NA – Experiment 10 did not include a protein pull-down component and instead tested cytoplasmic proteins processed using an in-gel sample preparation.

Table 6.3. Proteins pulled down with caC beads (Experiment 1).

Name	UniProt ID	Biological Process	Number of Unique Peptides
Regulator of nonsense transcripts 1	Q92900	DNA replication	2
Polymeric immunoglobulin receptor	P01833	immune response	4
Far upstream element-binding protein 2	Q92945	mRNA/RNA splicing	22
Heat shock 70 kDa protein 1-like	P34931	cellular homeostasis	2
Cleavage stimulation factor subunit 2	P33240	mRNA/RNA splicing	5
Lupus la protein	P05455	tRNA metabolic process	5
Splicing factor 3B subunit 4	Q15427	DNA replication	2
Ig gamma-1 chain C region	P01857	immune response	3
Transcriptional activator protein pur-alpha	Q00577	transcriptional regulation	3
Zinc-alpha-2-clycoprotein	P25311	immune response	6
Cellular nucleic acid-binding protein	P62633	transcriptional regulation	2
Putative lipocalin 1-like protein 1	Q5VSP4	unknown	2
Embryonic stem cell-specific 5-hydroxymethylcytosine-binding protein	Q96FZ2	unknown	2

Further testing of the affinity pull-down methodology has focused on improving the specificity of proteins for epigenetic marks (Experiment 2, **Table 6.2**). Competitor DNA was added during the binding step to minimize non-specific binding. We employed synthetic poly dAdT and calf thymus DNA. The proteins non-specifically binding to any DNA would also bind to the competitor DNA. Our intent was to depleting general DNA binding proteins, allowing for proteins with specific affinity for epigenetic DNA marks to be detected. The resulting proteins were analyzed by gel electrophoresis and Coomassie staining, but we did not observe any protein bands on the gel (not shown). The lack of sufficient protein for staining indicated that we needed to scale up the nuclear extract input for the pull-down to have enough sample for the proteomics experiments.

The first trial attempt using all five DNA baits (C, MeC, hmC, fC, and caC) was performed with N = 1 using the methodology of Spruijt et al (Experiment 3, **Table 6.2**).[26] This experiment employed 400 µg of nuclear protein extract incubated with 4 nmol of each DNA for a period of 2 hours at 4 ℃. Proteins were eluted by heating at 90 ℃ in 100 mM NaCl for 10 min, and were further processed by filter aided sample preparation (FASP) (**Scheme 6.1**).[384] The FASP procedure allows protein desalting prior to the necessary protein quantitation, reduction, and alkylation steps prior to digestion. Each of the capture reactions yielded less than 50 µg of total protein, which was digested using trypsin (1 µg) overnight. The peptides are then quantified prior to TMT labeling. The peptides were labeled by scaling down the amount of TMT reagent in a total volume of 45 µL. The labeled peptides were combined in equal microgram amounts, which is the most common way of

combining peptide samples. These samples showed very poor labeling when analyzed. The remaining peptides were re-labeled in an attempt to get better yields and re-analyzed. The reanalysis showed only 28 proteins, which were found labeled in four out of the five tags. These poor labeling resulting from too little tag being used relative to the quantity of peptides being labeled. From subsequent experiments, we learned that labeling of peptides requires a final concentration of 30% ACN to be successful.

In our next experiment, we pull-down experiments were conducted in triplicate using synthetic DNA duplexes containing C, MeC, hmC, fC, and caC (**Table 6.1**). The pull-down experiment utilized 30 µL of beads, 4 nmol of DNA, and the amount of nuclear extract protein was increased to 800 µg to increase the quantity of protein pulled down by each DNA bait (Experiment 4, **Table 6.2**). These samples were processed using the FASP methodology and the peptides quantified. To facilitate protein quantification across multiple runs, a reference sample was created. The reference sample was composed of 0.4 µg of peptides taken from each FASP prepared sample (**Scheme 6.2**). TMT-sixplex labels were used to label 2.4 µg of peptides from each sample and the reference sample. These labeled peptides were analyzed using a nanoLC-MS^3 method on an Orbitrap Fusion mass spectrometer.[379] In this analysis, 371 proteins were identified, however only 81% of the peptides were labeled. After further experiments, we learned the insufficient peptide labeling was due to the need for a 30% final concentration of ACN during the labeling reaction.

TMT quantitation results were processed using Proteome Discoverer (Thermo Fisher) compiled into Venn diagram using a tool available from Bioinformatics and Evolutionary Genomics (http://bioinformatics.psb.ugent.be/webtools/Venn/) (**Figure 6.2A**). Each epigenetic mark-containing DNA strand was recognized by a similar number of protein readers (**Figure 6.2A**). The quantitative results were averaged over the three replicates to generate a heat map of protein affinity (**Figure 6.3**). The heat map shows a range of binding affinities for each protein, from the most preferred epigenetic marks (in yellow) to the least preferential modification (in blue). The relative quantitation data were also graphed as log plots which report the fold-change binding preference for one modification over another on the x-axis and another pair of modifications on the y-axis (**Figure 6.4**). A full list of proteins specifically exhibiting different binding affinities are detailed in **Appendix A5**.

The proteins identified in the pull-downs as having an increased affinity toward specific epigenetic marks were analyzed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) to identify any enrichment for biological processes.[382] From the input proteins, DAVID identified the biological processes associated with each protein. The biological processes which had an enrichment p-value less than 0.05 were graphed into pie charts to show the relative contribution of each biological process (**Figure 6.5**). The proteins which preferentially bound to hmC showed an increase in the biological processes for DNA damage repair, transcription, and the immune response. Proteins pulled-down by fC containing DNA were enriched for the biological processes associated

with DNA repair, DNA metabolic processes, and cellular response to stress. caC containing DNA baits pulled-down proteins associated with DNA repair processes, RNA processing, and response to stress (**Figure 6.5**).

Scheme 6.2. TMT labeling scheme and the use of a reference sample for relative quantification. Samples from three replicates are assigned mass tags (126, 127, 128, 129, or 130). Preparation of the reference sample, right side, is accomplished by combining equal amounts of each sample and labeling with the same tag (131). This reference sample was included in each group of samples to allow for relative quantitation across all samples.

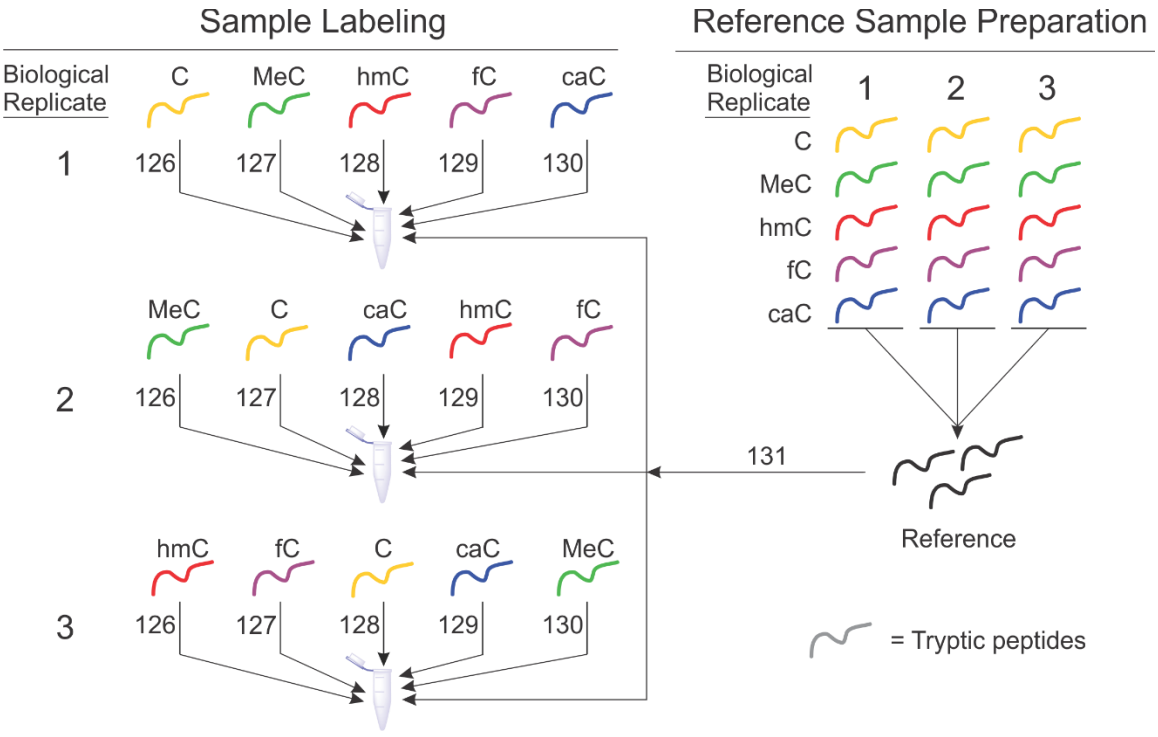


Figure 6.2. Proteins identified in affinity pull down experiments with DNA containing specific epigenetic marks. **A.** Venn diagram of protein binding showing protein numbers for each group. **B.** Total numbers of proteins found bound to each DNA duplex.

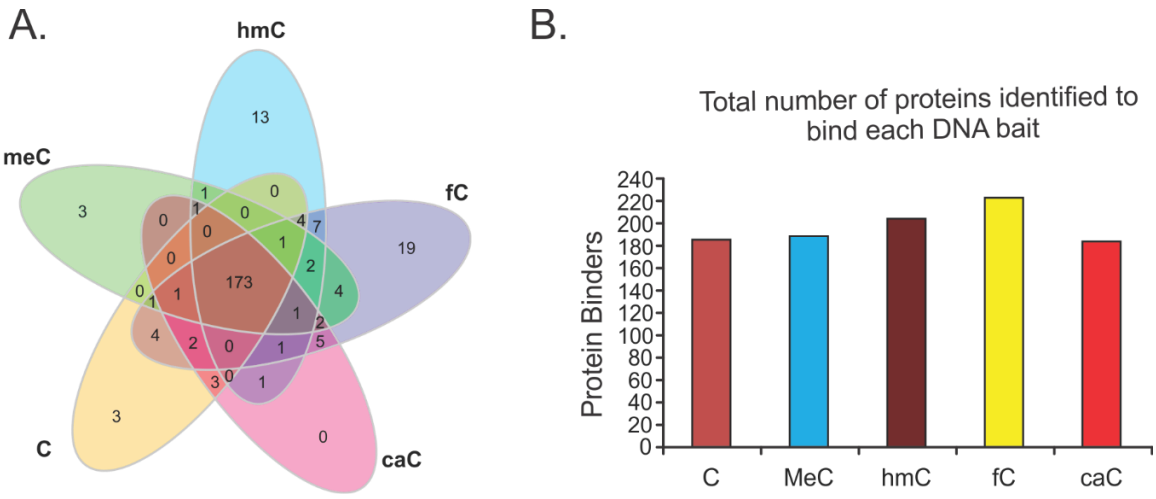


Figure 6.3. Heatmap of proteins pulled down with each DNA probe and analyzed using nanoLC-NSI-MS^3 methodology using TMT tags for quantification. Three replicates were averaged to a single value normalized using Cluster 3.0. The quantitation was visualized using Java TreeView. Gray spaces are samples in which the protein was not quantified.

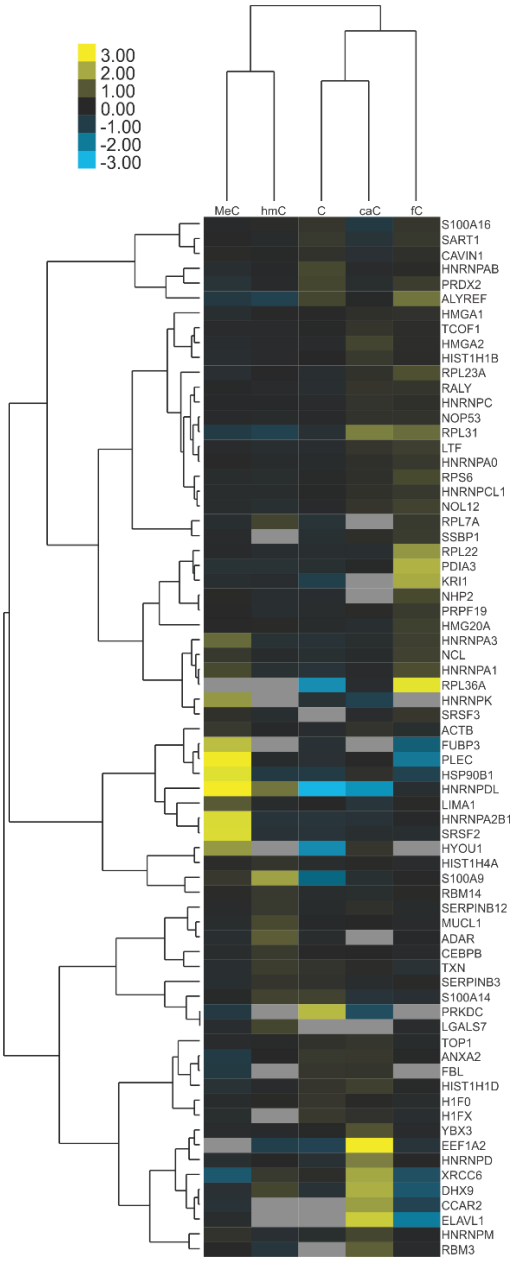


Figure 6.4. Relative protein affinity towards each epigenetic mark defined by log plots. In these plots, an increase in affinity for a specific modification is observed in movement to the right on the x-axis, or up on the y-axis. **A.** Affinity toward MeC and hmC. **B.** Affinity toward MeC and fC. **C.** Affinity toward MeC and caC. **D.** Affinity toward fC and caC. **E.** Affinity toward hmC and fC. **F.** Affinity toward hmC and caC.

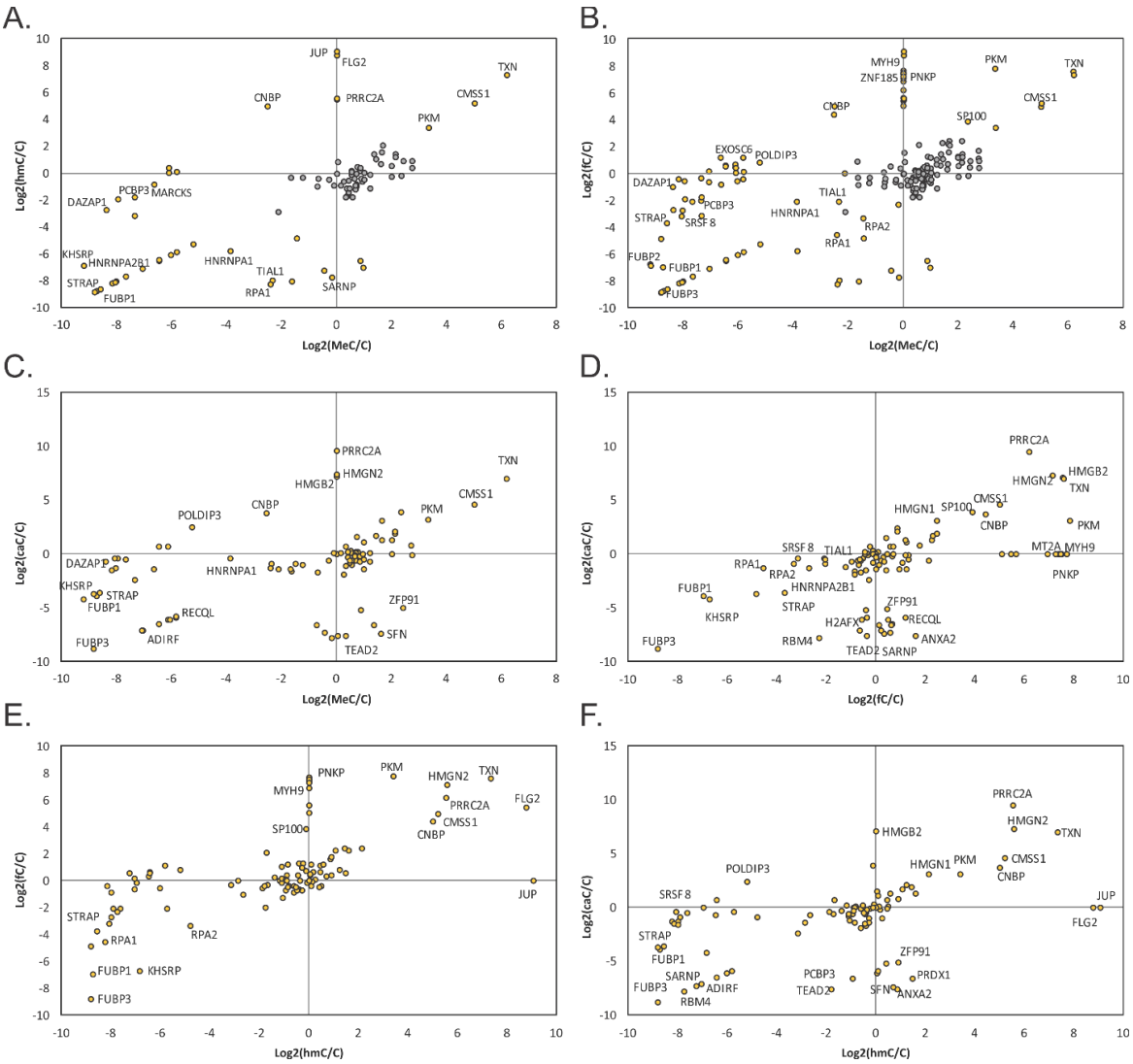
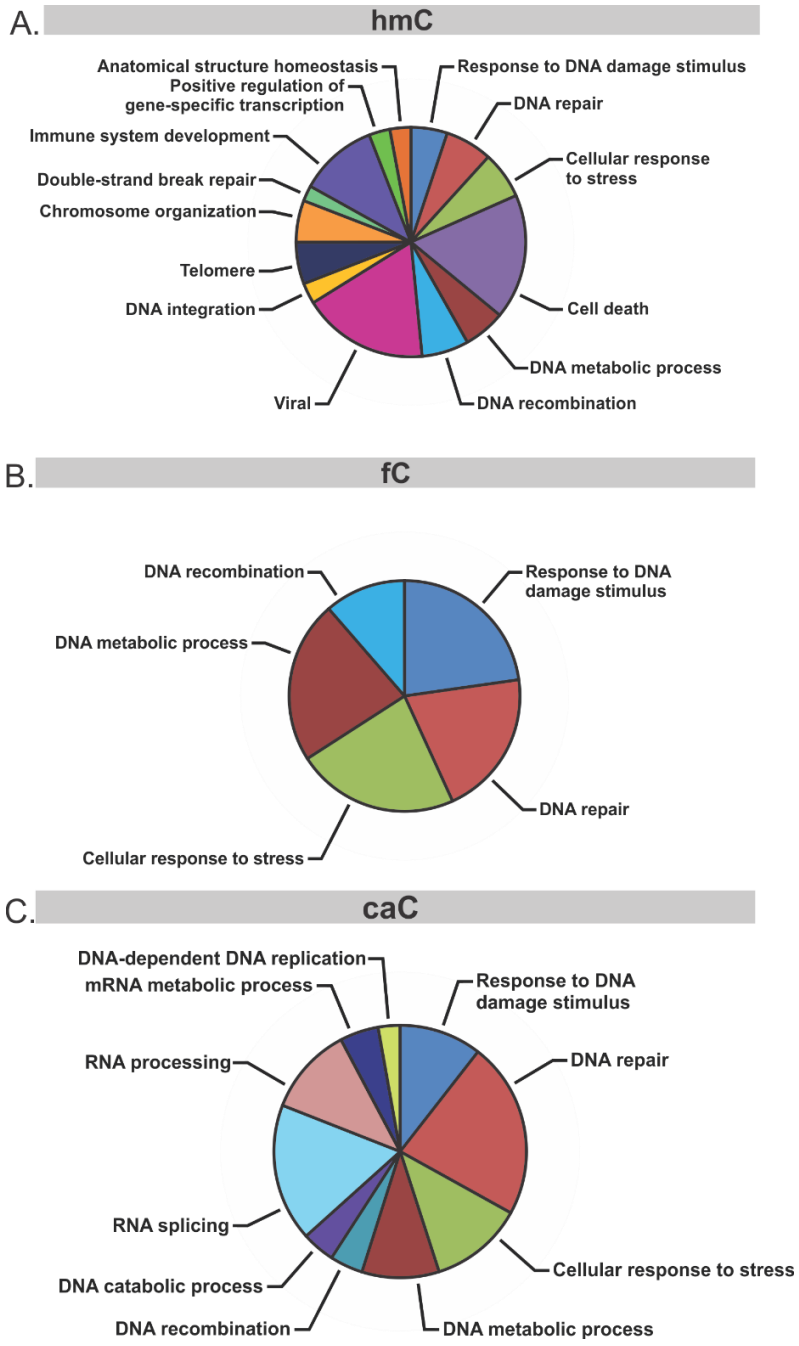


Figure 6.5. Biological processes associated with readers of oxiMeC as determined by the DAVID analysis tools. **A.** Readers of hmC containing DNA. **B.** readers of fC containing DNA. **C.** Readers of caC containing DNA.

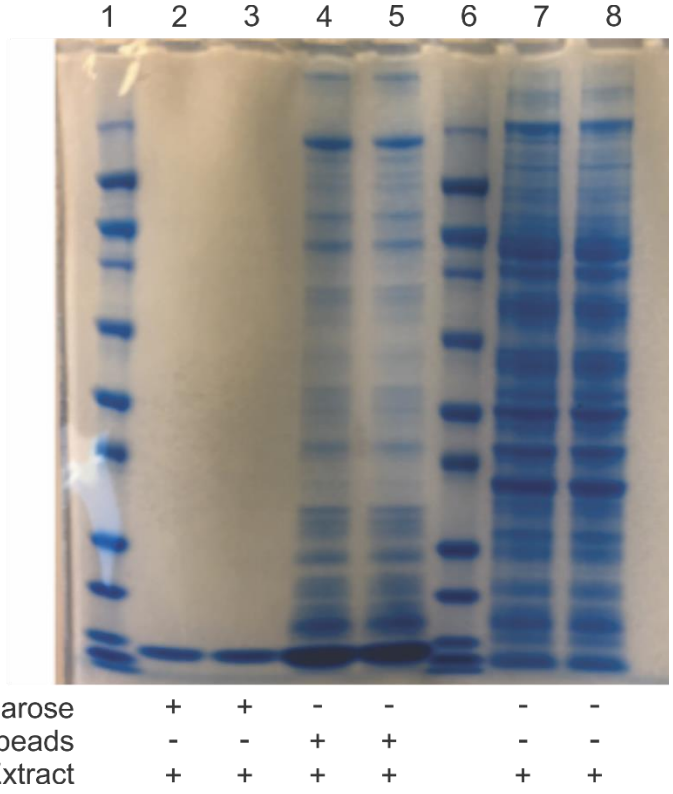


In Experiment 5, the same experiment was repeated in triplicate using a freshly prepared nuclear extract (Experiment 5, **Table 6.2**). The pull-downs were conducted using 30 µL of beads, 5 nmol of DNA, and 800 µg of nuclear extract. The proteins captured were processed with the FASP procedure with some modifications. Specifically, the FASP filters were passivated with a solution of 5% Tween 20 overnight to reduce protein loss during sample preparation. The buffer from the bead elution was exchanged with 8 M urea, 0.2% deoxycholic acid (DCA), 100 mM ammonium bicarbonate pH 8. Digestion was carried out in the presence of 0.2% DCA with trypsin. The DCA was added to prevent sticking of peptides to plastics during the processing and was removed during the stop and go extraction procedure (STAGE) for desalting peptides. The peptides were quantified and labeled as in Experiment 4, including the reference sample. These peptides were fractionated using basic reverse phase HPLC, and the eluted peptides were concatenated to 3 samples. When analyzed on an Orbitrap Fusion mass spectrometer, there were very few peptides found. From these experiments we learned that we needed to minimize the sample handling steps to reduce sample loss. Additionally, same day preparation of a nuclear extract is possible, but increases the difficulty of getting good samples with the lengthy pull-down procedure.

A follow-up experiment was carried out in which 4 of the DNA baits (C, MeC, hmC, and fC) were employed in triplicate at 2 nmol each bound to magnetic beads and 200 µg of protein (Experiment 6, **Table 6.2**). Captured and eluted proteins were prepared using the FASP methodology as described in the methods section. Following reduction,

alkylation, and digestion, there were very few peptides that moved through the filter membrane, and a second digestion was performed. Even after two successive digestions, a limited number of peptides were found in the triplicate samples. These were combined to afford 4 μg for each sample. Peptides were labeled with 4 μL of the reconstituted TMT reagent in a final 30% ACN and then STAGE tipped to desalt and remove extra reagents. These samples did show effective labeling efficiency (over 97%) indicating that the 30% ACN using in the labeling reaction is necessary to facilitate the reaction. Following the experiment, we directly compared the streptavidin coated magnetic beads (Thermo Fisher) to sepharose coated beads (GE Healthcare) in terms of non-specific binding. Using the magnetic beads, we found significant non-specific binding to the beads themselves and larger amounts of streptavidin released with our elution conditions as compared to streptavidin coated sepharose (**Figure 6.6**). The increased non-specific binding and greater leaching of streptavidin from magnetic beads during the elution indicates that sepharose beads are a better option for affinity pull downs.

Figure 6.6. Gel electrophoresis of proteins captured on different types of beads without DNA. 800 µg of nuclear extract for human bronchial epithelial cells was incubated with magnetic (Dynabeads, Thermo Fisher) or sepharose streptavidin coated beads (GE Healthcare), washed, and eluted with 100 mM NaCl. The eluted proteins were run out on the gel and stained with Coomassie stain. Lanes 1 and 6 are a protein MW ladder. Proteins pulled-down using streptavidin coated sepharose beads is in lanes 2 and 3. Proteins which bound to magnetic Dynabeads were run in lanes 4 and 5. Lanes 7 and 8 contain the whole nuclear protein extract.



We next explored the use of gel-based sample preparation methods for proteins captured by incubation with DNA baits as reported by Shevchenko et al.[385] Three synthetic DNA duplexes containing C, MeC, and hmC, were chosen for their selectivity in binding known proteins, specifically Tet1 for MeC and hmC and Uhrf2 for hmC. Three nanomoles of DNA was incubated with 900 µg of nuclear proteins using the methodology detailed in the methods section above (Experiment 7, **Table 6.2**). Proteins were eluted from the beads using 1X LDS, and each sample was split into three parts for further analysis. The first part was processed for western blotting against Tet1 and Uhrf2 to look for them in the pull-down solutions. The western blot was unable to identify either protein. Another portion of the protein extract was run into a gel for 10 minutes to desalt the proteins. The gel was processed using an adaption of Shevchenko et al.[385] The sections were cut from the gel and diced into 1 mm cubes. The gel cubes were washed using a solution of 25 mM ammonium bicarbonate and acetonitrile (1:1) and dehydrated with 100% ACN for 30 s. The gel pieces were rehydrated with 10 mM dithiothreitol in 20 mM ammonium bicarbonate and incubated at 56 ℃ for 1 h to reduce disulfide bonds. The supernatant was removed, 55 mM iodoacetamide in 25 mM ammonium bicarbonate was added, and then incubated for 30 min at room temperature. Following alkylation, the gel pieces were dehydrated and then rehydrated in 25 mM ammonium bicarbonate containing 2 µg of trypsin. Proteins were digested overnight at 37 ℃. Peptides were extracted by two successive incubations of the gel pieces with 50% ACN with 0.3% formic acid for 15 min each. Another extraction of the gel pieces was performed using 80% ACN with 0.3% formic acid. The extracts were

combined and dried to remove the formic acid. These peptides were labeled with 10 µL of TMT-0 reagent to test labeling efficiency in a single sample using the methodology described above. Excess reagent and salts were removed using C18 spin tips according the manufacturer's direction as described above. The remaining sample was run out on a gel and cut into 5 MW regions (more than 160 kDa, 80 – 160 kDa, 40 – 80 kDa, 20 – 40 kDa, and less than 20 kDa), digested using the procedure outlined above, and desalted using the C18 spin columns. Because of extensive mass spectrometry time needed to analyze a large number of samples arising from gel-based fractionation, only the unfractionated MeC, hmC, and the five fractions from the hmC gel were analyzed. These samples were analyzed using the Orbitrap Fusion mass spectrometer in an LC-MS/MS method. In this experiment, we were able to find a large number of proteins: 288 proteins in MeC DNA pull downs, 146 proteins in hmC DNA pull downs, and 288 proteins across the 5 fractions of hmC pull downs. Importantly, proteins identified in the gel-fractionated sample pulled down by hmC included key proteins of interest including Tet1, Uhrf2, Uhrf1, and Thy28 (**Table 6.4**). This indicates that fractionation of samples is necessary to facilitate detection of low abundance proteins.

Table 6.4. Unique peptides of key protein readers of hmC containing DNA identified via nanoHPLC-ESI^{+}-MS/MS.

Protein (MW)	Unique peptides (fractionation)	Unique Peptides (without fractionation)
Uhrf1 (91 kDa)	4	0
Thy28 (26 kDa)	10	5
RBM14 (69 kDa)	16	10
RPL26 (17 kDa)	1	2
PNKP (57 kDa)	8	0
Uhrf2 (90 kDa)	1	0

Following the promising detection of Tet1 and Uhrf2 peptides in the hmC pull-down, the same pull-down was repeated in triplicate in order to examine the reproducibility of the method (Experiment 8, **Table 6.2**). Sections of the gel were labeled with different TMT-sixplex tags and analyzed using the Orbitrap Fusion mass spectrometer. Following TMT labeling, samples were analyzed by nanoHPLC-NSI-MS2 on Orbitrap Fusion. However, we were unable to identify the majority of proteins identified from the previous fractionation experiment (Experiment 7), and there was poor reproducibility across three replicates using gel fractionation. Significant loss of key proteins after TMT-labeling was concerning. We investigated the loss of proteins by labeling the fraction from the previous experiment which had indicated the presence of Tet1 and Uhrf2. The TMT-labeled peptides are more hydrophobic and their respective retention times were shifted much later in the LC-MS/MS gradient. A possible cause of the loss of these key signals due to the later retention of all peptides is the co-elution of more abundant peptides at these later times.

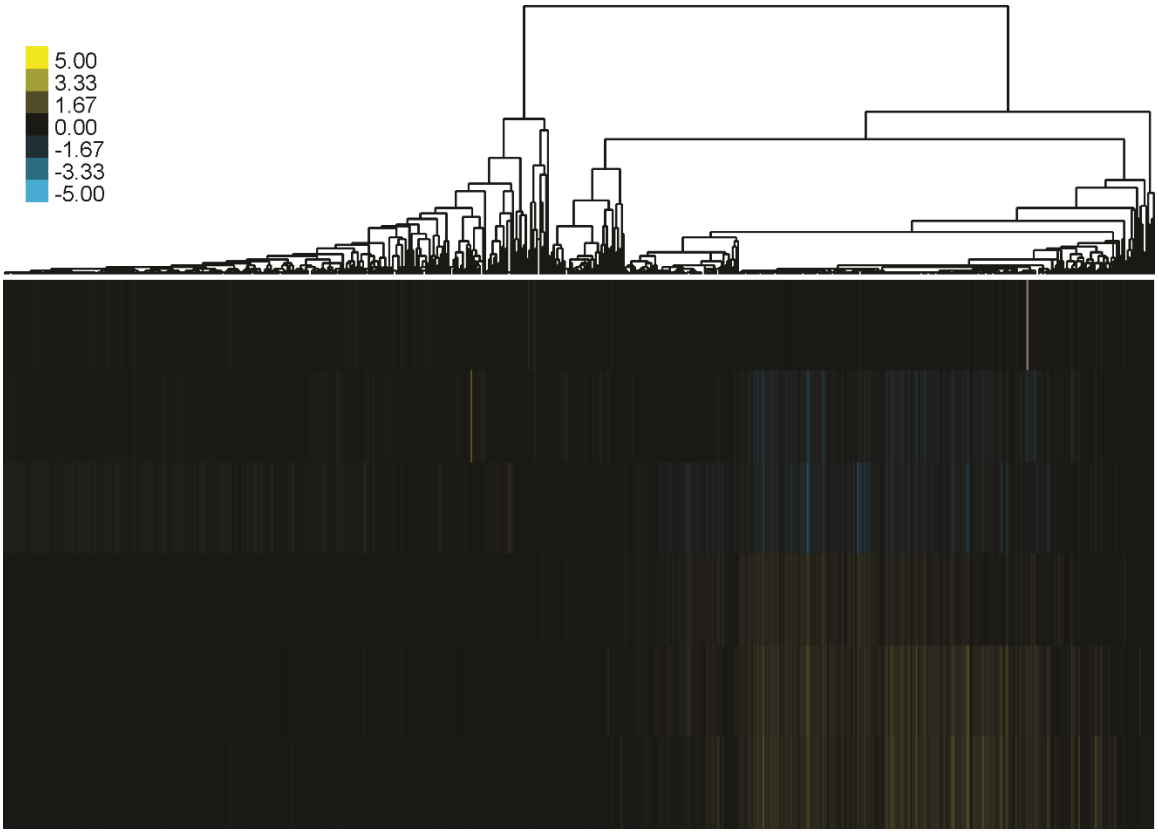
After the repeated poor-quality samples resulting from using the methodology described by Spruijt et al.,[26] we decided to test the competing method from Iularo et al. (Experiment 9, **Table 6.2**).[26-27] Iularo et al. bound DNA to the beads overnight at 4 °C.[27] The pull-down with nuclear protein extract was carried out for 15 min.[27] In addition to significantly different timing used for the experiment, Iularo et al. used different buffers for each step.[27] Using cytosine containing DNA in duplicate, the two sets of samples were processed according to their respective procedures. Proteins were eluted using 0.1% SDS and heating at 90 °C for 10 min. The eluted proteins were processed using FASP, the

peptides desalted with C18 spin columns, labeled with TMT-sixplex, combined, and cleaned-up using a second C18 spin column. The samples were analyzed on an Orbitrap Fusion mass spectrometer using the LC-MS/MS methodology described above. After the sample preparation, the LC-MS/MS runs showed no peptides in the samples. This is not consistent with results of an earlier experiment utilizing the same nuclear extract (Experiment 8), in which proteins were identified following analysis. It is unclear at this time what prevented the successful preparation of these samples.

Following the poor samples from the prior experiment, we set out to test the ability of the methodology to successfully prepare samples that had been digested in gel, which was thus far the most promising experimental methodology thus far (Experiment 10, **Table 6.2**). Cytosolic protein extract produced while preparing the nuclear extract was run into a gel in 6 separate lanes for 10 min. The proteins were digested using the in-gel digestion described above. The peptides were cleaned up using reverse phase SPE on Strata-X polymeric reverse phase cartridges. SPE were conditioned using 1 mL of acetonitrile then 1 mL of water, each with 0.1% formic acid. The samples were loaded, washed with 1 mL of 0.1% formic acid, and eluted with 0.1% formic acid in acetonitrile. An equal amount of peptides, 2.5 µg, was labeled with TMT-sixplex reagent as described above. The labeled peptides were combined, and desalted and fractionated with the basic reverse phase fractionation kit. The 8 elutions for TMT reagents were concatenated to 4 fractions. The four fractions were analyzed on an Orbitrap Fusion mass spectrometer using the nanoLC-MS/MS method described above. The data was analyzed using Proteome Discoverer 2.1

and the relative amount of protein in each sample quantified. These quantifier ion abundancies were normalized and clustered using Cluster 3.0 and visualized in a heatmap using Java TreeView (**Figure 6.7**). In this experiment, 1600 proteins were identified, with labeling efficiency of 99.3%. This experiment shows that the steps for processing samples after the pull-down are working and labeling with TMT reagent can be successfully carried out with a low microgram amount of peptides.

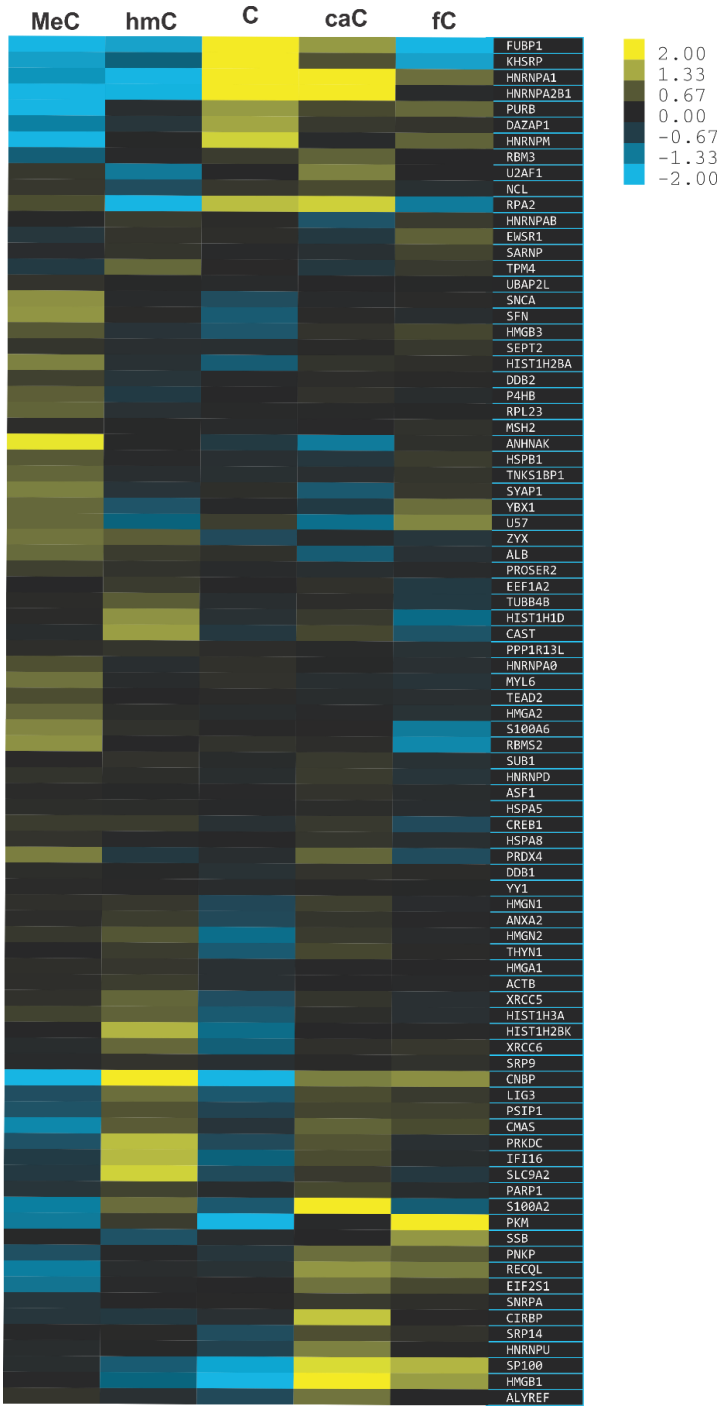
Figure 6.7. Heatmap of protein quantitation from cytoplasmic proteins using in-gel digestion, C18 clean-up, TMT labeling, and basic reverse phase fractionation. The protein quantitation is normalized to the mean for each protein in the channel.



Overall, the results of Experiment 10 indicate that the proteomics part of the assay is working, however our procedure may not pull down sufficient amounts of low abundance proteins such as known binders of oxiMeC such as Tet1 and Uhrf2. The next aspect of the pull-down tested focused on the DNA baits (Experiment 11, **Table 6.2**). Up to this point, DNA baits were prepared by solid phase synthesis on a DNA synthesizer and required significant efforts to purify and quantify them. We drew inspiration from the Iularo et al. work, which employed PCR to produce DNA baits.[27] Our DNA baits were made using a biotinylated primer and focused on the promoter region of the Rassf1 gene (Sequence: 5′-GGAAGGAGCTGAGGAGAGCCGCGCAATGGAAACCTGGGTGCAGGGACTGTGGGGCCCGAAGGCGGGGCTGGGCGCGCTCTCGCAGAGCCCCCCCCGCCTTGCCCTTCCTTCCCTCCTTCGTCCCCTCCTCACACCCCACCCCGGACGGCCACAACGACGGCGACCGCAAAGCACCACGCGGAGATACCCGTGTTTCTGGAGGCCAGCTTTAC-3′). We employed 50 ng of DNA for each pull-down replicate which was bound overnight at 4 ºC. The bound DNA was incubated with 75 µg of protein for 15 min, and the beads were washed 6 times (0.2 mM EDTA, 20% Glycerol, 20 mM Hepes-KOH pH 7.9, 0.1 M KCl, 1 mM DTT, 1 mM protease inhibitor PMSF, 0.1% Triton X-100).[27] The proteins were eluted with 0.1% SDS and processed using standard FASP methodology. Peptides were cleaned-up with C18 spin columns, TMT tagged, combined, and desalted and fractionated using the basic reverse phase fractionation kit. The TMT specific fractions were concatenated to 4 fractions and analyzed by nanoLC-MS/MS. Using the new capture

system, 86 proteins were identified, which is much less than expected, and TMT labeling was 96%. The quantitative results were clustered using Cluster 3.0 and visualized using Java TreeView (**Figure 6.8**). From this figure we can see that although 86 proteins were identified, there are many gaps in the quantitative data, suggesting that insufficient protein amounts were captured.

Figure 6.8. Heatmap of proteins identified and quantified in a protein pull-down using 212 BP DNA duplexes representing promoter region of the Rassf1 gene synthesized by PCR.



Several additional observations were made that could have a potentially significant impact in the methodology. While the published protocols[384] suggest washing the FASP filters with 0.1% FA in ACN, we found that these conditions can split the plastic filter devices along the seam. This was observed one time in 13 out of the 17 filter devices used, which had to be replaced with new Amicon 3k filters. Washing with 50:50 water:ACN with 0.1% FA avoids these issues. During the basic reverse phase fractionation kit sample preparation, each plastic tube was washed with 100 µL of 0.1% FA in can with vortexing in order to rinse away easily dissolved plastics and polymers from the tube surface. Thus far this appears to work, although a direct comparison has not been made at this time.

6.4 Conclusions and future directions

Our efforts directed toward a quantitative proteomics methodology for protein readers of oxiMeC in the lung has not yet yielded a finalized, reliable method for identifying epigenetic readers in the lung. In the process of developing this methodology, many aspects of the affinity pull-down have been explored and refined. In the areas of TMT labeling, sample processing, and data analysis we have identified effective processes. To achieve good labeling of peptides with TMT reagent requires the reaction to have 30% acetonitrile during the reaction. Many tests with in-gel digestion and FASP have been utilized. At this time, the high variability of peptide extraction from the gel at the levels present in these samples precludes the use of gel-based method. FASP has been more successful than gels in terms of peptide recovery for downstream processing and is recommended for future experiments. Although the complexity of pulled-down proteins is significantly less than a total proteome from whole tissue or nuclear extract, significantly better data has been collected using the Pierce™ high pH reversed-phase peptide fractionation kit and concatenation to a smaller number of fractions. This kit has the benefit of both desalting the sample, a necessary step, and orthogonally fractionating the peptides. Advances in the software for analyzing protein samples, especially the newer versions of Proteome Discoverer (version 2.1 and later) have allowed more complex quantitation calculations including using a reference sample to compare across multiple runs. Using programs such as Cluster 3.0, Java TreeView, and IPA improve understanding of the data with respect to reproducibility for Cluster 3.0 and Java TreeView and biological

significance with IPA. The application of multinotch MS^3 methodology has thus far proven detrimental to collecting good data due to the signal loss.[379] Instead, in our hands MS^2 based methodology provided better protein IDs.

Future work for this project broadly includes completing an effective pull-down with reproducible quantitation and identifying of protein readers of all four epigenetic marks. The next steps in completing this goal are to produce biotinylated DNA strands in large quantities using PCR or solid phase DNA synthesis. Data reproducibility is the second challenge and is reliant on the consistency of the technique used to prepare samples prior to combining after labeling. This can be achieved by reducing unnecessary steps prior to TMT tagging, or by changing the quantitative method altogether, for example by using label-free proteomics or by introducing stable isotope labeling of amino acids in cell culture (SILAC). Finally, identifying new readers of oxiMeC requires a secondary assay to validate those proteins such as electrophoretic mobility shift assays or western blotting. An optimized pulldown protocol taking into accounts all of our previous tests in given in Appendix A4.

VII. CONCLUSIONS

Epigenetic regulation describes changes in gene expression that do not result from changes in the gene sequence.[1] The ability of cells to maintain precise control over tissue specific gene expression patterns and to modify them as needed is critical for normal cellular development and homeostasis.[11, 386] Epigenetic regulation is achieved through reversible modification of DNA, RNA, and histones.[1] The primary epigenetic modification of DNA is methylation of cytosine at the C5 position to generate MeC.[11] Methylation on DNA is introduced via *de novo* methyltransferases DNMT3a/b and maintained through cell division by maintenance methyltransferase DNMT1.[11] MeC can be oxidized by TET to hmC, fC, and caC, a process known to induce DNA demethylation and gene reactivation (**Figure 1.2**).[13-14, 16, 21] Enrichment of oxiMeC at specific gene elements identified by DNA sequencing and proteomics studies support the idea that each oxidized form of MeC may have its own epigenetic function.[26-27, 205-206, 209-210] The precise balance of DNA methylation and demethylation can be influenced by chemical exposures to exogenous agents such as inflammatory agents present in cigarette smoke.

In Chapter II of this thesis, a kinetic assay of DNMT1 mediated methylation was developed to investigate the effects of extended C5-alkyl substituents such as ethyl, vinyl, and propyl groups on DNMT1 maintenance methylase activity (**Scheme 2.1**). We utilized synthetic DNA duplexes containing site-specific modification containing the extended C5-alkyl substituents (**Table 2.2**) to evaluate the enzymatic activity of DNMT1. We employed SPE enrichment coupled with HPLC-ESI$^+$-MS/MS to quantify the kinetics of methyl

transfer by isotope dilution tandem mass spectrometry (**Scheme 2.6**). Using this methodology, we found that DNMT1 was capable of methylating DNA containing the native substrate (MeC) and 5-ethyl-dC (V_{max} = 9.6 x 10^{-2} nM/min vs. 2.4 x 10^{-2} nM/min, respectively), but not 5-vinyl-dC or 5-propyl-dC (**Figure 2.6**). We created a homology model of the human DNMT1 protein and employed molecular mechanics calculations to assess the effects of increasing C5-alkyl side chain on the target recognition domain of the protein. From these analyses, we found that both 5-vinyl-dC and 5-propyl-dC disrupt the interactions of the target recognition domain of the enzyme with DNA, preventing the formation of a catalytically active conformation required for methyl transfer. This is consistent with the results from previous work showing that large side chains such 5-propyne-dC interfered with DNMT1 acitivity.[19]

In Chapter III of this thesis, we utilized the kinetic assay developed in Chapter II to investigate the kinetics of DNMT1 maintenance methylation toward DNA duplexes containing oxidized forms of MeC (**Table 3.2**). We employed offline HPLC clean-up coupled with the HPLC-ESI^+-MS/MS developed in Chapter II to determine the kinetics of DNMT1 activity (**Scheme 3.2**). We observed that DNMT1 was able to methylate DNA containing MeC, hmC, and fC with a reduced efficiency as the C5-methyl group was oxidized (V_{max}/K_m = 6.7, 2.1, and 0.85, x10^{-2} $min^{-1}M^{-1}$, **Table 3.3**). We employed homology model developed in Chapter II to carry out molecular dynamics simulations of DNA bound to the protein. Using this model, we observed changes in DNA and protein structure over 100 ns and found that hydrophilic nature of oxiMeC interferes with

hydrophobic interactions between TRD and MeC. These findings support the idea that formation of oxiMeC contributes to passive DNA demethylation by hindering the activity of DNMT1.

In Chapters IV and V of this thesis, we investigated epigenetic changes in the A/J mouse model of smoking induced lung cancer. Smoking induced lung tumors are characterized by profound epigenetic changes such as aberrant patterns of DNA methylation and hydroxymethylation, deregulated histone acetylation, and altered gene expression levels.[4, 107-108] These epigenetic alterations are likely to represent an important event in lung tumor development.[108] However, the origins and the timing of smoking-mediated epigenetic dysregulation in the lung remain largely unknown. DNA isolated form lung tissues of from A/J mice exposed to cigarette smoke for 10 weeks was subjected to RRBS and oxoRRBS analyses to determine site-specific changes in MeC and hmC.[140] We found that cigarette smoke exposure altered site specific methylation and hydroxymethylation patterns of many cancer related genes (**Figure 4.2**). To identify ETS components responsible for these effects, A/J mice were exposed to NNK, LPS, or NNK and LPS for 2, 6, or 27 weeks. Lung DNA was isolated and used to quantify the global levels of MeC, hmC, and fC. To facilitate quantitation, we synthesized the necessary stable isotope labeled internal standards and employed an HPLC enrichment, which was coupled to a novel HPLC-ESI$^+$-MS/MS method developed in our laboratory (**Figure 4.1 and 4.3**). In mice treated for 6 weeks with LPS, we observed a 0.2% increase in MeC, and a decrease in hmC (**Figure 4.6**), while the amounts of fC were increased (**Figure 4.6**). In the 6-week

treatment we also observed a decrease in global histone acetylation at H3K14 and H3K23 with LPS treatment (**Figure 4.11**). To determine whether these epigenetic changes observed in whole lung were representative of the type II cells of the lung, we treated mice with LPS along and isolated the type II lung cells by panning (Chapter V). Histone acetylation analysis of the type II cells revealed acetylation changes that matched those observed in whole lung (**Figure 5.5**). The RNA from type II cells was subjected to RNA-Seq analysis to identify changes transcription (**Figure 5.3**). The transcriptional changes observed are in agreement with previous work (**Figure 5.6**).[142] Correlating DNA methylation and hydroxymethylation changes observed in mice exposed to ETS to gene expression changed in LPS-treated mice using IPA showed an enrichment in cancer related genes (**Tables 4.3 – 4.6**). These results provide initial evidence that cigarette smoke alters site-specific methylation and hydroxymethylation of genes that are associated with the cancer phenotype. In addition, 100 most deregulated methylation and hydroxymethylation sites were correlated with gene expression from type II cells to look for enrichment (**Figures 5.7 and 5.8**). Increases in MeC and hmC were enriched for genes with decreasing expression.

In Chapter VI of this thesis, we set out to develop an affinity proteomics methodology to identify the lung nuclear protein readers that preferentially bind oxidized forms of MeC in DNA. To identify these epigenetic readers of epigenetic marks in DNA, we designed dsDNA containing site-specific C, MeC, hmC, fC, and caC (**Table 6.1**). Multiple conditions were tested to facilitate the identification of protein readers (**Table**

6.2). While the downstream sample processing steps (TMT labeling, sample cleanup, and mass spectrometry analysis) have been well refined to give consistent results (**Figure 6.7**), effective affinity capture remained elusive. This chapter details the development of a method for carrying out the affinity proteomics including: the use of MS^2 based methodology, efficient peptide tagging, sample clean-up, and fractionation.

In summary, the research detailed in this Thesis provides novel information about the kinetics of DNA methylation and possible demethylation mechanisms (Chapters II and III), epigenetic changes associated with smoking and inflammation (Chapters IV and V), and the identities of protein readers of various epigenetic marks in human lung (Chapter VI). These findings pave the way for future work to identify epigenetic mechanisms of cancer and to develop novel epigenetic modulators as future therapeutic agents.

VIII. FUTURE DIRECTIONS

8.1 Integrated 'omics approach to investigate epigenetic changes in type II pulmonary cells following exposure to cigarette smoke

As described in chapter IV, lung cancer is responsible for over 30% of cancer related deaths worldwide, and cigarette smoking is a major contributor to the prevalence of lung cancer.[101] Epigenetic deregulation in likely to play a major role in lung cancer etiology. Our previous animal studies utilized whole lung tissue comprised of a diverse mixture of cells. Within these experiments, the epigenetic changes in a single cell type may be masked by the other cells, or immune cell infiltration. Additionally, the effects of cigarette smoke exposure on type II lung cells, which are expected to be sites of lung cancer initiation, has not been systematically studied.[225] By using flow cytometry to isolate type II cells in mice exposed to cigarette smoke and single agents, it is possible to systematically study the effects of cigarette smoke on this key cell population. To study epigenetic changes in type II cells, laboratory A/J mice of both sexes can be exposed to cigarette smoke, or LPS to induce epigenetic changes. After isolation of the type II cells from the lung, DNA, RNA, and proteins can be extracted to facilitate a thorough investigation of epigenetic changes. The DNA can be used in global HPLC-ESI-MS/MS quantitation of epigenetic changes as well as RRBS and oxoRRBS to determine site-specific changes of DNA methylation and hydroxymethylation. The RNA can be used for RNA-Seq analysis of the transcriptional changes. A global, quantitative proteomics approach can be used to investigate changes in protein abundance. Integrating genomics, transcriptomics, and

proteomics would provide a comprehensive analysis of the epigenetic changes and their outcomes occurring in type II cells following exposure to cigarette smoke.

8.2 Investigation of protein readers of MeC, hmC, fC, and caC in the lung and lung tumors

Although protein readers of MeC and its oxidized forms in the brain have been investigated,[26] little is known about the proteins interacting with these epigenetics marks in other tissue types. Proteomes of different cell and tissue types vary significantly, and the same mark can be recognized by a different set of readers. In Chapter VI of this thesis, we discussed the development of quantitative affinity proteomics methodology to identify nuclear protein readers in human bronchial epithelial cells. The current methodology has well established steps for protein and peptide processing following the affinity purification. Section 6.4 of Chapter VI lays out the immediate steps to progress the current state of the project such as efficient pull downs, minimization of steps, and validation of targets with secondary assays. Further developments in determining the readers of epigenetic modifications by quantitative proteomics could implement additional genomic regions, such as enhancers, as the region of choice can have an effect on the proteins bound.[27] Additionally, the use of recently discovered epigenetic modifications such as N6MedA[36] or 4mC as DNA targets would provide additional insight on the functions of those modifications in human lung. Protein pull-downs can be further expanded to explore the composition of readers in tumors induced by carcinogens and cigarette smoke.

8.3 Improved analytical methods for quantitation of epigenetic modifications of DNA, MeC, hmC, fC, and caC

Epigenetic modifications of DNA (MeC, hmC, fC, and caC) vary significantly in their cellular abundance, ranging from 5% of total Cs for MeC to 5 x 10^{-5} % of total Cs for caC.[13-14, 16] This creates a problem with instrument dynamic range and accurate quantification. To make things more challenging, these cytidine analogues are very polar molecules and do not retain well on typical reverse phase HPLC columns. Sample preparation requirements for these analytes has proven difficult due to requirements for time consuming sample enrichment by offline HPLC (see Chapter IV). Recently reported efforts to use a 2D-HPLC-ESI-MS/MS method caught our attention as a potential way to eliminate derivatization and pre-enrichment needed to analyze these modifications.[159] A similar methodology can be employed with reduced solvent flow rates to improve sensitivity and conform to the technical specifications of available equipment. Alternatively, the described 2D-HPLC-ESI-MS/MS method could be adapted to employ nanoLC for the second dimension and nanospray ionization for a greater increase in sensitivity.

8.4 Epigenetic changes following exposure to E-liquid vapor

The rise of electronic cigarettes as an alternative to smoking creates new public health risks.[387] The public perceives electronic cigarette use as safer than traditional smoking and this can contribute to freer use when normal cigarettes would be avoided,

such as during pregnancy.[387] The electronic cigarette liquid has been shown to cause DNA damage in the lungs, heart, and bladder of exposed mice.[388] Further, the use of electronic cigarettes has been shown to suppress immune and inflammatory response genes in nasal epithelial cells of smokers in a similar manner to smoking.[389] There are many varieties of E-liquids on the market. These E-liquids contain many different components in addition to nicotine, such as flavorants and humectants, which can create additional risks.[390-393] It would be interesting to identify epigenetic changes (global DNA epigenetic modifications, RRBS, RNA-Seq, and histone modifications) in the lung of A/J mice exposed to E-cig vapor. These experiments would reveal whether the use of electronic cigarettes induces epigenetic changes similar to those caused by traditional tobacco smoking.

BIBLIOGRAPHY

1. Allis, C. D.; Jenuwein, T., The molecular hallmarks of epigenetic control. *Nat Rev Genet* **2016,** *17* (8), 487-500.
2. Chapman, V. L.; Terranova, R.; Moggs, J. G.; Kimber, I.; Dearman, R. J., Evaluation of 5-methylcytosine and 5-hydroxymethylcytosine as potential biomarkers for characterisation of chemical allergens. *Toxicology* **2016,** *340*, 17-26.
3. Sharma, S.; Kelly, T. K.; Jones, P. A., Epigenetics in cancer. *Carcinogenesis* **2010,** *31* (1), 27-36.
4. Biswas, S.; Thomas, A. A.; Chakrabarti, S., LncRNAs: Proverbial Genomic "Junk" or Key Epigenetic Regulators During Cardiac Fibrosis in Diabetes? *Front Cardiovasc Med* **2018,** *5*, 28.
5. Wu, X.; Zhang, Y., TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet* **2017**.
6. International Human Genome Sequencing, C., Finishing the euchromatic sequence of the human genome. *Nature* **2004,** *431* (7011), 931-45.
7. Bestor, T. H., The DNA methyltransferases of mammals. *Hum Mol Genet* **2000,** *9* (16), 2395-402.
8. Warnecke, P. M.; Bestor, T. H., Cytosine methylation and human cancer. *Curr. Opin. Oncol* **2000,** *12* (1), 68-73.
9. Panning, B.; Jaenisch, R., RNA and the epigenetic regulation of X chromosome inactivation. *Cell* **1998,** *93* (3), 305-8.
10. Song, J.; Teplova, M.; Ishibe-Murakami, S.; Patel, D. J., Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* **2012,** *335* (6069), 709-12.
11. Klose, R. J.; Bird, A. P., Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* **2006,** *31* (2), 89-97.
12. Chedin, F.; Lieber, M. R.; Hsieh, C. L., The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc Natl Acad Sci U S A* **2002,** *99* (26), 16916-21.
13. Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L.; Rao, A., Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **2009,** *324* (5929), 930-935.
14. Pfaffeneder, T.; Hackner, B.; Truss, M.; Munzel, M.; Muller, M.; Deiml, C. A.; Hagemeier, C.; Carell, T., The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed Engl* **2011,** *50* (31), 7008-7012.
15. Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **2011,** *333* (6047), 1300-1303.
16. He, Y. F.; Li, B. Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L.; Sun, Y.; Li, X.; Dai, Q.; Song, C. X.; Zhang, K.; He, C.; Xu, G. L., Tet-mediated

formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **2011,** *333* (6047), 1303-1307.
17. Hu, L.; Lu, J.; Cheng, J.; Rao, Q.; Li, Z.; Hou, H.; Lou, Z.; Zhang, L.; Li, W.; Gong, W.; Liu, M.; Sun, C.; Yin, X.; Li, J.; Tan, X.; Wang, P.; Wang, Y.; Fang, D.; Cui, Q.; Yang, P.; He, C.; Jiang, H.; Luo, C.; Xu, Y., Structural insight into substrate preference for TET-mediated oxidation. *Nature* **2015,** *527* (7576), 118-122.
18. Hu, L.; Li, Z.; Cheng, J.; Rao, Q.; Gong, W.; Liu, M.; Shi, Y. G.; Zhu, J.; Wang, P.; Xu, Y., Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **2013,** *155* (7), 1545-1555.
19. Valinluck, V.; Sowers, L. C., Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res* **2007,** *67* (3), 946-950.
20. Kotandeniya, D.; Seiler, C. L.; Fernandez, J.; Pujari, S. S.; Curwick, L.; Murphy, K.; Wickramaratne, S.; Yan, S.; Murphy, D.; Sham, Y. Y.; Tretyakova, N. Y., Can 5-methylcytosine analogues with extended alkyl side chains guide DNA methylation? *Chem Commun (Camb)* **2018,** *54* (9), 1061-1064.
21. Kweon, S. M.; Zhu, B.; Chen, Y.; Aravind, L.; Xu, S. Y.; Feldman, D. E., Erasure of Tet-Oxidized 5-Methylcytosine by a SRAP Nuclease. *Cell Rep* **2017,** *21* (2), 482-494.
22. Song, C. X.; He, C., Potential functional roles of DNA demethylation intermediates. *Trends Biochem. Sci* **2013,** *38* (10), 480-484.
23. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Burgess, H. E.; Iurlaro, M.; Reik, W.; Murrell, A.; Balasubramanian, S., 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol* **2015**.
24. Kriukiene, E.; Liutkeviciute, Z.; Klimasauskas, S., 5-Hydroxymethylcytosine--the elusive epigenetic mark in mammalian DNA. *Chem. Soc. Rev* **2012,** *41* (21), 6916-6930.
25. Fu, Y.; He, C., Nucleic acid modifications with epigenetic significance. *Curr. Opin. Chem. Biol* **2012,** *16* (5-6), 516-524.
26. Spruijt, C. G.; Gnerlich, F.; Smits, A. H.; Pfaffeneder, T.; Jansen, P. W.; Bauer, C.; Munzel, M.; Wagner, M.; Muller, M.; Khan, F.; Eberl, H. C.; Mensinga, A.; Brinkman, A. B.; Lephikov, K.; Muller, U.; Walter, J.; Boelens, R.; van Ingen, H.; Leonhardt, H.; Carell, T.; Vermeulen, M., Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **2013,** *152* (5), 1146-1159.
27. Iurlaro, M.; Ficz, G.; Oxley, D.; Raiber, E. A.; Bachman, M.; Booth, M. J.; Andrews, S.; Balasubramanian, S.; Reik, W., A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **2013,** *14* (10), R119.
28. Ngo, T. T.; Yoo, J.; Dai, Q.; Zhang, Q.; He, C.; Aksimentiev, A.; Ha, T., Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat Commun* **2016,** *7*, 10813.
29. Skvortsova, K.; Zotenko, E.; Luu, P. L.; Gould, C. M.; Nair, S. S.; Clark, S. J.; Stirzaker, C., Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin* **2017,** *10*, 16.

30. Globisch, D.; Munzel, M.; Muller, M.; Michalakis, S.; Wagner, M.; Koch, S.; Bruckl, T.; Biel, M.; Carell, T., Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS. One* **2010,** *5* (12), e15367.
31. Gackowski, D.; Zarakowska, E.; Starczak, M.; Modrzejewska, M.; Olinski, R., Tissue-Specific Differences in DNA Modifications (5-Hydroxymethylcytosine, 5-Formylcytosine, 5-Carboxylcytosine and 5-Hydroxymethyluracil) and Their Interrelationships. *PLoS One* **2015,** *10* (12), e0144859.
32. Heyn, H.; Esteller, M., An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell* **2015,** *161* (4), 710-3.
33. Koziol, M. J.; Bradshaw, C. R.; Allen, G. E.; Costa, A. S. H.; Frezza, C.; Gurdon, J. B., Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat Struct Mol Biol* **2016,** *23* (1), 24-30.
34. Achwal, C. W.; Iyer, C. A.; Chandra, H. S., Immunochemical evidence for the presence of 5mC, 6mA and 7mG in human, Drosophila and mealybug DNA. *FEBS Lett* **1983,** *158* (2), 353-8.
35. Ratel, D.; Ravanat, J. L.; Charles, M. P.; Platet, N.; Breuillaud, L.; Lunardi, J.; Berger, F.; Wion, D., Undetectable levels of N6-methyl adenine in mouse DNA: Cloning and analysis of PRED28, a gene coding for a putative mammalian DNA adenine methyltransferase. *FEBS Lett* **2006,** *580* (13), 3179-84.
36. Wu, T. P.; Wang, T.; Seetin, M. G.; Lai, Y.; Zhu, S.; Lin, K.; Liu, Y.; Byrum, S. D.; Mackintosh, S. G.; Zhong, M.; Tackett, A.; Wang, G.; Hon, L. S.; Fang, G.; Swenberg, J. A.; Xiao, A. Z., DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **2016,** *532* (7599), 329-33.
37. Yao, B.; Cheng, Y.; Wang, Z.; Li, Y.; Chen, L.; Huang, L.; Zhang, W.; Chen, D.; Wu, H.; Tang, B.; Jin, P., DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nat Commun* **2017,** *8* (1), 1122.
38. Ehrlich, M.; Wilson, G. G.; Kuo, K. C.; Gehrke, C. W., N4-methylcytosine as a minor base in bacterial DNA. *J Bacteriol* **1987,** *169* (3), 939-43.
39. Timinskas, A.; Butkus, V.; Janulaitis, A., Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. Classification of all DNA methyltransferases. *Gene* **1995,** *157* (1-2), 3-11.
40. Yu, M.; Ji, L.; Neumann, D. A.; Chung, D. H.; Groom, J.; Westpheling, J.; He, C.; Schmitz, R. J., Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. *Nucleic Acids Res* **2015,** *43* (21), e148.
41. Schubeler, D., Function and information content of DNA methylation. *Nature* **2015,** *517* (7534), 321-6.
42. Williams, K.; Christensen, J.; Pedersen, M. T.; Johansen, J. V.; Cloos, P. A.; Rappsilber, J.; Helin, K., TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **2011,** *473* (7347), 343-8.
43. Raiber, E. A.; Beraldi, D.; Ficz, G.; Burgess, H. E.; Branco, M. R.; Murat, P.; Oxley, D.; Booth, M. J.; Reik, W.; Balasubramanian, S., Genome-wide distribution of 5-

formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol* **2012,** *13* (8), R69.
44. Neri, F.; Incarnato, D.; Krepelova, A.; Rapelli, S.; Anselmi, F.; Parlato, C.; Medana, C.; Dal Bello, F.; Oliviero, S., Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Rep* **2015**.
45. Kellinger, M. W.; Song, C. X.; Chong, J.; Lu, X. Y.; He, C.; Wang, D., 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol* **2012,** *19* (8), 831-833.
46. Josling, G. A.; Selvarajah, S. A.; Petter, M.; Duffy, M. F., The role of bromodomain proteins in regulating gene expression. *Genes (Basel)* **2012,** *3* (2), 320-43.
47. Allfrey, V. G.; Faulkner, R.; Mirsky, A. E., Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci U S A* **1964,** *51*, 786-94.
48. Bannister, A. J.; Kouzarides, T., Regulation of chromatin by histone modifications. *Cell Res* **2011,** *21* (3), 381-95.
49. Grunstein, M., Histone acetylation in chromatin structure and transcription. *Nature* **1997,** *389* (6649), 349-52.
50. Wei, Y.; Liang, J.; Zhang, R.; Guo, Y.; Shen, S.; Su, L.; Lin, X.; Moran, S.; Helland, A.; Bjaanaes, M. M.; Karlsson, A.; Planck, M.; Esteller, M.; Fleischer, T.; Staaf, J.; Zhao, Y.; Chen, F.; Christiani, D. C., Epigenetic modifications in KDM lysine demethylases associate with survival of early-stage NSCLC. *Clin Epigenetics* **2018,** *10*, 41.
51. Yuan, H.; Marmorstein, R., Histone acetyltransferases: Rising ancient counterparts to protein kinases. *Biopolymers* **2013,** *99* (2), 98-111.
52. Bernstein, B. E.; Kamal, M.; Lindblad-Toh, K.; Bekiranov, S.; Bailey, D. K.; Huebert, D. J.; McMahon, S.; Karlsson, E. K.; Kulbokas, E. J., 3rd; Gingeras, T. R.; Schreiber, S. L.; Lander, E. S., Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **2005,** *120* (2), 169-81.
53. Yang, X. J.; Seto, E., HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene* **2007,** *26* (37), 5310-8.
54. Yang, X. J.; Seto, E., The Rpd3/Hda1 family of lysine deacetylases: from bacteria and yeast to mice and men. *Nat Rev Mol Cell Biol* **2008,** *9* (3), 206-18.
55. Collins, R. E.; Tachibana, M.; Tamaru, H.; Smith, K. M.; Jia, D.; Zhang, X.; Selker, E. U.; Shinkai, Y.; Cheng, X., In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases. *J Biol Chem* **2005,** *280* (7), 5563-70.
56. Cheng, X.; Collins, R. E.; Zhang, X., Structural and sequence motifs of protein (histone) methylation enzymes. *Annu Rev Biophys Biomol Struct* **2005,** *34*, 267-94.
57. Dong, X.; Weng, Z., The correlation between histone modifications and gene expression. *Epigenomics* **2013,** *5* (2), 113-6.
58. Barski, A.; Cuddapah, S.; Cui, K.; Roh, T. Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K., High-resolution profiling of histone methylations in the human genome. *Cell* **2007,** *129* (4), 823-37.

59. Bernstein, B. E.; Mikkelsen, T. S.; Xie, X.; Kamal, M.; Huebert, D. J.; Cuff, J.; Fry, B.; Meissner, A.; Wernig, M.; Plath, K.; Jaenisch, R.; Wagschal, A.; Feil, R.; Schreiber, S. L.; Lander, E. S., A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **2006,** *125* (2), 315-26.
60. Schlesinger, Y.; Straussman, R.; Keshet, I.; Farkash, S.; Hecht, M.; Zimmerman, J.; Eden, E.; Yakhini, Z.; Ben-Shushan, E.; Reubinoff, B. E.; Bergman, Y.; Simon, I.; Cedar, H., Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **2007,** *39* (2), 232-6.
61. Shi, Y.; Lan, F.; Matson, C.; Mulligan, P.; Whetstine, J. R.; Cole, P. A.; Casero, R. A.; Shi, Y., Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **2004,** *119* (7), 941-53.
62. You, A.; Tong, J. K.; Grozinger, C. M.; Schreiber, S. L., CoREST is an integral component of the CoREST- human histone deacetylase complex. *Proc Natl Acad Sci U S A* **2001,** *98* (4), 1454-8.
63. Zheng, Y. C.; Ma, J.; Wang, Z.; Li, J.; Jiang, B.; Zhou, W.; Shi, X.; Wang, X.; Zhao, W.; Liu, H. M., A Systematic Review of Histone Lysine-Specific Demethylase 1 and Its Inhibitors. *Med Res Rev* **2015,** *35* (5), 1032-71.
64. Kong, X.; Ouyang, S.; Liang, Z.; Lu, J.; Chen, L.; Shen, B.; Li, D.; Zheng, M.; Li, K. K.; Luo, C.; Jiang, H., Catalytic mechanism investigation of lysine-specific demethylase 1 (LSD1): a computational study. *PLoS One* **2011,** *6* (9), e25444.
65. Culhane, J. C.; Szewczuk, L. M.; Liu, X.; Da, G.; Marmorstein, R.; Cole, P. A., A mechanism-based inactivator for histone demethylase LSD1. *J Am Chem Soc* **2006,** *128* (14), 4536-7.
66. Whetstine, J. R.; Nottke, A.; Lan, F.; Huarte, M.; Smolikov, S.; Chen, Z.; Spooner, E.; Li, E.; Zhang, G.; Colaiacovo, M.; Shi, Y., Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell* **2006,** *125* (3), 467-81.
67. Morris, K. V.; Mattick, J. S., The rise of regulatory RNA. *Nat Rev Genet* **2014,** *15* (6), 423-37.
68. Roundtree, I. A.; He, C., RNA epigenetics--chemical messages for posttranscriptional gene regulation. *Curr Opin Chem Biol* **2016,** *30*, 46-51.
69. Boccaletto, P.; Machnicka, M. A.; Purta, E.; Piatkowski, P.; Baginski, B.; Wirecki, T. K.; de Crecy-Lagard, V.; Ross, R.; Limbach, P. A.; Kotter, A.; Helm, M.; Bujnicki, J. M., MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* **2018,** *46* (D1), D303-D307.
70. Williams, T. M.; Selegue, J. E.; Werner, T.; Gompel, N.; Kopp, A.; Carroll, S. B., The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila. *Cell* **2008,** *134* (4), 610-23.
71. Ambros, V., The functions of animal microRNAs. *Nature* **2004,** *431* (7006), 350-5.
72. Heimberg, A. M.; Sempere, L. F.; Moy, V. N.; Donoghue, P. C.; Peterson, K. J., MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* **2008,** *105* (8), 2946-50.

73. Park, C. Y.; Choi, Y. S.; McManus, M. T., Analysis of microRNA knockouts in mice. *Hum Mol Genet* **2010,** *19* (R2), R169-75.
74. Leonardo, T. R.; Schultheisz, H. L.; Loring, J. F.; Laurent, L. C., The functions of microRNAs in pluripotency and reprogramming. *Nat Cell Biol* **2012,** *14* (11), 1114-21.
75. Bracken, C. P.; Gregory, P. A.; Khew-Goodall, Y.; Goodall, G. J., The role of microRNAs in metastasis and epithelial-mesenchymal transition. *Cell Mol Life Sci* **2009,** *66* (10), 1682-99.
76. Holoch, D.; Moazed, D., RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* **2015,** *16* (2), 71-84.
77. Kopp, F.; Mendell, J. T., Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **2018,** *172* (3), 393-407.
78. Schmitz, K. M.; Mayer, C.; Postepska, A.; Grummt, I., Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **2010,** *24* (20), 2264-9.
79. Rinn, J. L.; Chang, H. Y., Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **2012,** *81*, 145-66.
80. Khalil, A. M.; Guttman, M.; Huarte, M.; Garber, M.; Raj, A.; Rivea Morales, D.; Thomas, K.; Presser, A.; Bernstein, B. E.; van Oudenaarden, A.; Regev, A.; Lander, E. S.; Rinn, J. L., Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **2009,** *106* (28), 11667-72.
81. Geisler, S.; Coller, J., RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol* **2013,** *14* (11), 699-712.
82. Zhubi, A.; Chen, Y.; Dong, E.; Cook, E. H.; Guidotti, A.; Grayson, D. R., Increased binding of MeCP2 to the GAD1 and RELN promoters may be mediated by an enrichment of 5-hmC in autism spectrum disorder (ASD) cerebellum. *Transl. Psychiatry* **2014,** *4*, e349.
83. Kadowaki, H.; Nishitoh, H.; Urano, F.; Sadamitsu, C.; Matsuzawa, A.; Takeda, K.; Masutani, H.; Yodoi, J.; Urano, Y.; Nagano, T.; Ichijo, H., Amyloid β induces neuronal cell death through ROS-mediated ASK1 activation. *Cell Death. Differ* **2005,** *12* (1), 19-24.
84. Laird, P. W., Cancer epigenetics. *Hum Mol Genet* **2005,** *14 Spec No 1*, R65-76.
85. Gnyszka, A.; Jastrzebski, Z.; Flis, S., DNA methyltransferase inhibitors and their emerging role in epigenetic therapy of cancer. *Anticancer Res* **2013,** *33* (8), 2989-96.
86. Tsai, H. C.; Li, H.; Van Neste, L.; Cai, Y.; Robert, C.; Rassool, F. V.; Shin, J. J.; Harbom, K. M.; Beaty, R.; Pappou, E.; Harris, J.; Yen, R. W.; Ahuja, N.; Brock, M. V.; Stearns, V.; Feller-Kopman, D.; Yarmus, L. B.; Lin, Y. C.; Welm, A. L.; Issa, J. P.; Minn, I.; Matsui, W.; Jang, Y. Y.; Sharkis, S. J.; Baylin, S. B.; Zahnow, C. A., Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell* **2012,** *21* (3), 430-46.
87. Flis, S.; Gnyszka, A.; Flis, K., DNA methyltransferase inhibitors improve the effect of chemotherapeutic agents in SW48 and HT-29 colorectal cancer cells. *PLoS One* **2014,** *9* (3), e92305.

88. Ruiz-Magana, M. J.; Rodriguez-Vargas, J. M.; Morales, J. C.; Saldivia, M. A.; Schulze-Osthoff, K.; Ruiz-Ruiz, C., The DNA methyltransferase inhibitors zebularine and decitabine induce mitochondria-mediated apoptosis and DNA damage in p53 mutant leukemic T cells. *Int J Cancer* **2012,** *130* (5), 1195-207.
89. Kantarjian, H.; Issa, J. P.; Rosenfeld, C. S.; Bennett, J. M.; Albitar, M.; DiPersio, J.; Klimek, V.; Slack, J.; de Castro, C.; Ravandi, F.; Helmer, R., 3rd; Shen, L.; Nimer, S. D.; Leavitt, R.; Raza, A.; Saba, H., Decitabine improves patient outcomes in myelodysplastic syndromes: results of a phase III randomized study. *Cancer* **2006,** *106* (8), 1794-803.
90. Zhao, M.; Wang, J.; Liao, W.; Li, D.; Li, M.; Wu, H.; Zhang, Y.; Gershwin, M. E.; Lu, Q., Increased 5-hydroxymethylcytosine in CD4(+) T cells in systemic lupus erythematosus. *J Autoimmun* **2016,** *69*, 64-73.
91. Somineni, H. K.; Zhang, X.; Biagini Myers, J. M.; Kovacic, M. B.; Ulm, A.; Jurcak, N.; Ryan, P. H.; Khurana Hershey, G. K.; Ji, H., Ten-eleven translocation 1 (TET1) methylation is associated with childhood asthma and traffic-related air pollution. *J Allergy Clin Immunol* **2016,** *137* (3), 797-805 e5.
92. Pfeifer, G. P.; Xiong, W.; Hahn, M. A.; Jin, S. G., The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res* **2014,** *356* (3), 631-641.
93. Lee, J. H.; Mahendran, A.; Yao, Y.; Ngo, L.; Venta-Perez, G.; Choy, M. L.; Kim, N.; Ham, W. S.; Breslow, R.; Marks, P. A., Development of a histone deacetylase 6 inhibitor and its biological effects. *Proc Natl Acad Sci U S A* **2013,** *110* (39), 15704-9.
94. Laubach, J. P.; Moreau, P.; San-Miguel, J. F.; Richardson, P. G., Panobinostat for the Treatment of Multiple Myeloma. *Clin Cancer Res* **2015,** *21* (21), 4767-73.
95. Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; Philpott, M.; Munro, S.; McKeown, M. R.; Wang, Y.; Christie, A. L.; West, N.; Cameron, M. J.; Schwartz, B.; Heightman, T. D.; La Thangue, N.; French, C. A.; Wiest, O.; Kung, A. L.; Knapp, S.; Bradner, J. E., Selective inhibition of BET bromodomains. *Nature* **2010,** *468* (7327), 1067-73.
96. Ardekani, A. M.; Naeini, M. M., The Role of MicroRNAs in Human Diseases. *Avicenna J Med Biotechnol* **2010,** *2* (4), 161-79.
97. Soifer, H. S.; Rossi, J. J.; Saetrom, P., MicroRNAs in disease and potential therapeutic applications. *Mol Ther* **2007,** *15* (12), 2070-9.
98. Poliseno, L.; Salmena, L.; Zhang, J.; Carver, B.; Haveman, W. J.; Pandolfi, P. P., A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **2010,** *465* (7301), 1033-8.
99. Shen, J.; Samul, R.; Silva, R. L.; Akiyama, H.; Liu, H.; Saishin, Y.; Hackett, S. F.; Zinnen, S.; Kossen, K.; Fosnaugh, K.; Vargeese, C.; Gomez, A.; Bouhana, K.; Aitchison, R.; Pavco, P.; Campochiaro, P. A., Suppression of ocular neovascularization with siRNA targeting VEGF receptor 1. *Gene Ther* **2006,** *13* (3), 225-34.
100. Bitko, V.; Musiyenko, A.; Shulyayeva, O.; Barik, S., Inhibition of respiratory viruses by nasally administered siRNA. *Nat Med* **2005,** *11* (1), 50-5.

101. Siegel, R. L.; Miller, K. D.; Jemal, A., Cancer statistics, 2018. *CA Cancer J Clin* **2018,** *68* (1), 7-30.
102. Hecht, S. S.; Carmella, S. G.; Murphy, S. E.; Akerkar, S.; Brunnemann, K. D.; Hoffmann, D., A tobacco-specific lung carcinogen in the urine of men exposed to cigarette smoke. *N. Engl. J. Med* **1993,** *329* (21), 1543-1546.
103. Tretyakova, N. Y.; Chiang, S. Y.; Walker, V. E.; Swenberg, J. A., Quantitative analysis of 1,3-butadiene-induced DNA adducts in vivo and in vitro using liquid chromatography electrospray ionization tandem mass spectrometry. *J Mass Spectrom* **1998,** *33* (4), 363-376.
104. Ronai, Z. A.; Gradia, S.; Peterson, L. A.; Hecht, S. S., G to A transitions and G to T transversions in codon 12 of the *Ki-ras* oncogene isolated from mouse lung tumors induced by 4- (methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and related DNA methylating and pyridyloxobutylating agents. *Carcinogenesis* **1993,** *14* (11), 2419-2422.
105. Hasday, J. D.; Bascom, R.; Costa, J. J.; Fitzgerald, T.; Dubin, W., Bacterial endotoxin is an active component of cigarette smoke. *Chest* **1999,** *115* (3), 829-835.
106. Matter, B.; Seiler, C. L.; Murphy, K.; Ming, X.; Zhao, J.; Lindgren, B.; Jones, R.; Tretyakova, N., Mapping Three Guanine Oxidation Products along DNA Following Exposure to Three Types of Reactive Oxygen Species. *Free Radic Biol Med* **2018**.
107. Esteller, M.; Corn, P. G.; Baylin, S. B.; Herman, J. G., A gene hypermethylation profile of human cancer. *Cancer Res* **2001,** *61* (8), 3225-3229.
108. Kerr, K. M.; Galler, J. S.; Hagen, J. A.; Laird, P. W.; Laird-Offringa, I. A., The role of DNA methylation in the development and progression of lung adenocarcinoma. *Dis. Markers* **2007,** *23* (1-2), 5-30.
109. Selamat, S. A.; Galler, J. S.; Joshi, A. D.; Fyfe, M. N.; Campan, M.; Siegmund, K. D.; Kerr, K. M.; Laird-Offringa, I. A., DNA methylation changes in atypical adenomatous hyperplasia, adenocarcinoma in situ, and lung adenocarcinoma. *PLoS One* **2011,** *6* (6), e21443.
110. Tsou, J. A.; Galler, J. S.; Siegmund, K. D.; Laird, P. W.; Turla, S.; Cozen, W.; Hagen, J. A.; Koss, M. N.; Laird-Offringa, I. A., Identification of a panel of sensitive and specific DNA methylation markers for lung adenocarcinoma. *Mol Cancer* **2007,** *6*, 70.
111. Chi, P.; Allis, C. D.; Wang, G. G., Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer* **2010,** *10* (7), 457-69.
112. Ougolkov, A. V.; Bilim, V. N.; Billadeau, D. D., Regulation of pancreatic tumor cell proliferation and chemoresistance by the histone methyltransferase enhancer of zeste homologue 2. *Clin Cancer Res* **2008,** *14* (21), 6790-6.
113. Cao, Q.; Yu, J.; Dhanasekaran, S. M.; Kim, J. H.; Mani, R. S.; Tomlins, S. A.; Mehra, R.; Laxman, B.; Cao, X.; Yu, J.; Kleer, C. G.; Varambally, S.; Chinnaiyan, A. M., Repression of E-cadherin by the polycomb group protein EZH2 in cancer. *Oncogene* **2008,** *27* (58), 7274-84.
114. Wu, H.; Chen, X.; Xiong, J.; Li, Y.; Li, H.; Ding, X.; Liu, S.; Chen, S.; Gao, S.; Zhu, B., Histone methyltransferase G9a contributes to H3K27 methylation in vivo. *Cell Res* **2011,** *21* (2), 365-7.

115. Chen, M. W.; Hua, K. T.; Kao, H. J.; Chi, C. C.; Wei, L. H.; Johansson, G.; Shiah, S. G.; Chen, P. S.; Jeng, Y. M.; Cheng, T. Y.; Lai, T. C.; Chang, J. S.; Jan, Y. H.; Chien, M. H.; Yang, C. J.; Huang, M. S.; Hsiao, M.; Kuo, M. L., H3K9 histone methyltransferase G9a promotes lung cancer invasion and metastasis by silencing the cell adhesion molecule Ep-CAM. *Cancer Res* **2010,** *70* (20), 7830-40.
116. Li, Y.; Seto, E., HDACs and HDAC Inhibitors in Cancer Development and Therapy. *Cold Spring Harb Perspect Med* **2016,** *6* (10).
117. Barber, M. F.; Michishita-Kioi, E.; Xi, Y.; Tasselli, L.; Kioi, M.; Moqtaderi, Z.; Tennen, R. I.; Paredes, S.; Young, N. L.; Chen, K.; Struhl, K.; Garcia, B. A.; Gozani, O.; Li, W.; Chua, K. F., SIRT7 links H3K18 deacetylation to maintenance of oncogenic transformation. *Nature* **2012,** *487* (7405), 114-8.
118. Florczuk, M.; Szpechcinski, A.; Chorostowska-Wynimko, J., miRNAs as Biomarkers and Therapeutic Targets in Non-Small Cell Lung Cancer: Current Perspectives. *Target Oncol* **2017,** *12* (2), 179-200.
119. Fiori, M. E.; Villanova, L.; Barbini, C.; De Angelis, M. L.; De Maria, R., miR-663 sustains NSCLC by inhibiting mitochondrial outer membrane permeabilization (MOMP) through PUMA/BBC3 and BTG2. *Cell Death Dis* **2018,** *9* (2), 49.
120. Czubak, K.; Lewandowska, M. A.; Klonowska, K.; Roszkowski, K.; Kowalewski, J.; Figlerowicz, M.; Kozlowski, P., High copy number variation of cancer-related microRNA genes and frequent amplification of DICER1 and DROSHA in lung cancer. *Oncotarget* **2015,** *6* (27), 23399-416.
121. Alberg, A. J.; Brock, M. V.; Ford, J. G.; Samet, J. M.; Spivack, S. D., Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **2013,** *143* (5 Suppl), e1S-e29S.
122. Ansari, J.; Shackelford, R. E.; El-Osta, H., Epigenetics in non-small cell lung cancer: from basics to therapeutics. *Transl Lung Cancer Res* **2016,** *5* (2), 155-71.
123. Chu, B. F.; Karpenko, M. J.; Liu, Z.; Aimiuwu, J.; Villalona-Calero, M. A.; Chan, K. K.; Grever, M. R.; Otterson, G. A., Phase I study of 5-aza-2'-deoxycytidine in combination with valproic acid in non-small-cell lung cancer. *Cancer Chemother Pharmacol* **2013,** *71* (1), 115-21.
124. Lin, J.; Gilbert, J.; Rudek, M. A.; Zwiebel, J. A.; Gore, S.; Jiemjit, A.; Zhao, M.; Baker, S. D.; Ambinder, R. F.; Herman, J. G.; Donehower, R. C.; Carducci, M. A., A phase I dose-finding study of 5-azacytidine in combination with sodium phenylbutyrate in patients with refractory solid tumors. *Clin Cancer Res* **2009,** *15* (19), 6241-9.
125. Candelaria, M.; Gallardo-Rincon, D.; Arce, C.; Cetina, L.; Aguilar-Ponce, J. L.; Arrieta, O.; Gonzalez-Fierro, A.; Chavez-Blanco, A.; de la Cruz-Hernandez, E.; Camargo, M. F.; Trejo-Becerril, C.; Perez-Cardenas, E.; Perez-Plasencia, C.; Taja-Chayeb, L.; Wegman-Ostrosky, T.; Revilla-Vazquez, A.; Duenas-Gonzalez, A., A phase II study of epigenetic therapy with hydralazine and magnesium valproate to overcome chemotherapy resistance in refractory solid tumors. *Ann Oncol* **2007,** *18* (9), 1529-38.

126. Stathis, A.; Hotte, S. J.; Chen, E. X.; Hirte, H. W.; Oza, A. M.; Moretto, P.; Webster, S.; Laughlin, A.; Stayner, L. A.; McGill, S.; Wang, L.; Zhang, W. J.; Espinoza-Delgado, I.; Holleran, J. L.; Egorin, M. J.; Siu, L. L., Phase I study of decitabine in combination with vorinostat in patients with advanced solid tumors and non-Hodgkin's lymphomas. *Clin Cancer Res* **2011,** *17* (6), 1582-90.
127. Coussens, L. M.; Werb, Z., Inflammation and cancer. *Nature* **2002,** *420* (6917), 860-7.
128. Hanahan, D.; Weinberg, R. A., Hallmarks of cancer: the next generation. *Cell* **2011,** *144* (5), 646-74.
129. Colotta, F.; Allavena, P.; Sica, A.; Garlanda, C.; Mantovani, A., Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis* **2009,** *30* (7), 1073-81.
130. Straub, R. H.; Schradin, C., Chronic inflammatory systemic diseases: An evolutionary trade-off between acutely beneficial but chronically harmful programs. *Evol Med Public Health* **2016,** *2016* (1), 37-51.
131. Durham, A. L.; Adcock, I. M., The relationship between COPD and lung cancer. *Lung Cancer* **2015,** *90* (2), 121-7.
132. Freeman, H. J., Colorectal cancer risk in Crohn's disease. *World J Gastroenterol* **2008,** *14* (12), 1810-1.
133. Stinn, W.; Berges, A.; Meurrens, K.; Buettner, A.; Gebel, S.; Lichtner, R. B.; Janssens, K.; Veljkovic, E.; Xiang, Y.; Roemer, E.; Haussmann, H. J., Towards the validation of a lung tumorigenesis model with mainstream cigarette smoke inhalation using the A/J mouse. *Toxicology* **2013,** *305*, 49-64.
134. Mauderly, J. L.; Gigliotti, A. P.; Barr, E. B.; Bechtold, W. E.; Belinsky, S. A.; Hahn, F. F.; Hobbs, C. A.; March, T. H.; Seilkop, S. K.; Finch, G. L., Chronic inhalation exposure to mainstream cigarette smoke increases lung and nasal tumor incidence in rats. *Toxicol Sci* **2004,** *81* (2), 280-92.
135. Bogen, K. T.; Witschi, H., Lung tumors in A/J mice exposed to environmental tobacco smoke: estimated potency and implied human risk. *Carcinogenesis* **2002,** *23* (3), 511-9.
136. Curtin, G. M.; Higuchi, M. A.; Ayres, P. H.; Swauger, J. E.; Mosberg, A. T., Lung tumorigenicity in A/J and rasH2 transgenic mice following mainstream tobacco smoke inhalation. *Toxicol Sci* **2004,** *81* (1), 26-34.
137. Gordon, T.; Bosland, M., Strain-dependent differences in susceptibility to lung cancer in inbred mice exposed to mainstream cigarette smoke. *Cancer Lett* **2009,** *275* (2), 213-20.
138. Stinn, W.; Arts, J. H.; Buettner, A.; Duistermaat, E.; Janssens, K.; Kuper, C. F.; Haussmann, H. J., Murine lung tumor response after exposure to cigarette mainstream smoke or its particulate and gas/vapor phase fractions. *Toxicology* **2010,** *275* (1-3), 10-20.
139. Hutt, J. A.; Vuillemenot, B. R.; Barr, E. B.; Grimes, M. J.; Hahn, F. F.; Hobbs, C. H.; March, T. H.; Gigliotti, A. P.; Seilkop, S. K.; Finch, G. L.; Mauderly, J. L.; Belinsky, S. A., Life-span inhalation exposure to mainstream cigarette smoke induces lung cancer in

B6C3F1 mice through genetic and epigenetic pathways. *Carcinogenesis* **2005,** *26* (11), 1999-2009.
140. La Maestra, S.; D'Agostini, F.; Izzotti, A.; Micale, R. T.; Mastracci, L.; Camoirano, A.; Balansky, R.; Trosko, J. E.; Steele, V. E.; De Flora, S., Modulation by aspirin and naproxen of nucleotide alterations and tumors in the lung of mice exposed to environmental cigarette smoke since birth. *Carcinogenesis* **2015,** *36* (12), 1531-8.
141. Melkamu, T.; Qian, X.; Upadhyaya, P.; O'Sullivan, M. G.; Kassie, F., Lipopolysaccharide enhances mouse lung tumorigenesis: a model for inflammation-driven lung cancer. *Vet. Pathol* **2013,** *50* (5), 895-902.
142. Qian, X.; Khammanivong, A.; Song, J. M.; Teferi, F.; Upadhyaya, P.; Dickerson, E.; Kassie, F., RNA-sequencing studies identify genes differentially regulated during inflammation-driven lung tumorigenesis and targeted by chemopreventive agents. *Inflamm Res* **2015,** *64* (5), 343-61.
143. Takahashi, H.; Ogata, H.; Nishigaki, R.; Broide, D. H.; Karin, M., Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta- and JNK1-dependent inflammation. *Cancer Cell* **2010,** *17* (1), 89-97.
144. Hecht, S. S.; Isaacs, S.; Trushin, N., Lung tumor induction in A/J mice by the tobacco smoke carcinogens 4- (methylnitrosamino)-1-(3-pyridyl)-1-butanone and benzo[a]pyrene: a potentially useful model for evaluation of chemopreventive agents. *Carcinogenesis* **1994,** *15* (12), 2721-2725.
145. Hotchkiss, R. D., The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **1948,** *175* (1), 315-32.
146. Kriaucionis, S.; Heintz, N., The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **2009,** *324* (5929), 929-930.
147. Sano, H.; Royer, H. D.; Sager, R., Identification of 5-methylcytosine in DNA fragments immobilized on nitrocellulose paper. *Proc Natl Acad Sci U S A* **1980,** *77* (6), 3581-5.
148. Kurdyukov, S.; Bullock, M., DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)* **2016,** *5* (1).
149. Mizugaki, M.; Itoh, K.; Yamaguchi, T.; Ishiwata, S.; Hishinuma, T.; Nozaki, S.; Ishida, N., Preparation of a monoclonal antibody specific for 5-methyl-2'-deoxycytidine and its application for the detection of DNA methylation levels in human peripheral blood cells. *Biol Pharm Bull* **1996,** *19* (12), 1537-40.
150. Inoue, A.; Shen, L.; Dai, Q.; He, C.; Zhang, Y., Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **2011,** *21* (12), 1670-1676.
151. Valentini, E.; Zampieri, M.; Malavolta, M.; Bacalini, M. G.; Calabrese, R.; Guastafierro, T.; Reale, A.; Franceschi, C.; Hervonen, A.; Koller, B.; Bernhardt, J.; Slagboom, P. E.; Toussaint, O.; Sikora, E.; Gonos, E. S.; Breusing, N.; Grune, T.; Jansen, E.; Dolle, M. E.; Moreno-Villanueva, M.; Sindlinger, T.; Burkle, A.; Ciccarone, F.; Caiafa, P., Analysis of the machinery and intermediates of the 5hmC-mediated DNA

demethylation pathway in aging on samples from the MARK-AGE Study. *Aging (Albany NY)* **2016,** *8* (9), 1896-1922.
152. Habib, M.; Fares, F.; Bourgeois, C. A.; Bella, C.; Bernardino, J.; Hernandez-Blazquez, F.; de Capoa, A.; Niveleau, A., DNA global hypomethylation in EBV-transformed interphase nuclei. *Exp Cell Res* **1999,** *249* (1), 46-53.
153. Liu, M. Y.; DeNizio, J. E.; Kohli, R. M., Quantification of Oxidized 5-Methylcytosine Bases and TET Enzyme Activity. *Methods Enzymol* **2016,** *573*, 365-85.
154. Du, C.; Kurabe, N.; Matsushima, Y.; Suzuki, M.; Kahyo, T.; Ohnishi, I.; Tanioka, F.; Tajima, S.; Goto, M.; Yamada, H.; Tao, H.; Shinmura, K.; Konno, H.; Sugimura, H., Robust quantitative assessments of cytosine modifications and changes in the expressions of related enzymes in gastric cancer. *Gastric Cancer* **2015,** *18* (3), 516-25.
155. Bachman, M.; Uribe-Lewis, S.; Yang, X.; Williams, M.; Murrell, A.; Balasubramanian, S., 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* **2014,** *6* (12), 1049-55.
156. Pfaffeneder, T.; Spada, F.; Wagner, M.; Brandmayr, C.; Laube, S. K.; Eisen, D.; Truss, M.; Steinbacher, J.; Hackner, B.; Kotljarova, O.; Schuermann, D.; Michalakis, S.; Kosmatchev, O.; Schiesser, S.; Steigenberger, B.; Raddaoui, N.; Kashiwazaki, G.; Muller, U.; Spruijt, C. G.; Vermeulen, M.; Leonhardt, H.; Schar, P.; Muller, M.; Carell, T., Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **2014,** *10* (7), 574-81.
157. Tsuji, M.; Matsunaga, H.; Jinno, D.; Tsukamoto, H.; Suzuki, N.; Tomioka, Y., A validated quantitative liquid chromatography-tandem quadrupole mass spectrometry method for monitoring isotopologues to evaluate global modified cytosine ratios in genomic DNA. *J Chromatogr B Analyt Technol Biomed Life Sci* **2014,** *953-954*, 38-47.
158. Liu, S.; Wang, J.; Su, Y.; Guerrero, C.; Zeng, Y.; Mitra, D.; Brooks, P. J.; Fisher, D. E.; Song, H.; Wang, Y., Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic Acids Res* **2013,** *41* (13), 6421-6429.
159. Gackowski, D.; Starczak, M.; Zarakowska, E.; Modrzejewska, M.; Szpila, A.; Banaszkiewicz, Z.; Olinski, R., Accurate, Direct, and High-Throughput Analyses of a Broad Spectrum of Endogenously Generated DNA Base Modifications with Isotope-Dilution Two-Dimensional Ultraperformance Liquid Chromatography with Tandem Mass Spectrometry: Possible Clinical Implication. *Anal Chem* **2016,** *88* (24), 12128-12136.
160. Lin, X. C.; Zhang, T.; Liu, L.; Tang, H.; Yu, R. Q.; Jiang, J. H., Mass Spectrometry Based Ultrasensitive DNA Methylation Profiling Using Target Fragmentation Assay. *Anal Chem* **2016,** *88* (2), 1083-7.
161. Le, T.; Kim, K. P.; Fan, G.; Faull, K. F., A sensitive mass spectrometry method for simultaneous quantification of DNA methylation and hydroxymethylation levels in biological samples. *Anal Biochem* **2011,** *412* (2), 203-9.
162. Hayatsu, H.; Wataya, Y.; Kazushige, K., The addition of sodium bisulfite to uracil and to cytosine. *J Am Chem Soc* **1970,** *92* (3), 724-6.

163. Hayatsu, H., Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis--a personal account. *Proc Jpn Acad Ser B Phys Biol Sci* **2008,** *84* (8), 321-30.
164. Hayatsu, H.; Shiragami, M., Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry* **1979,** *18* (4), 632-7.
165. Clark, S. J.; Statham, A.; Stirzaker, C.; Molloy, P. L.; Frommer, M., DNA methylation: bisulphite modification and analysis. *Nat Protoc* **2006,** *1* (5), 2353-64.
166. Deng, J.; Shoemaker, R.; Xie, B.; Gore, A.; LeProust, E. M.; Antosiewicz-Bourget, J.; Egli, D.; Maherali, N.; Park, I. H.; Yu, J.; Daley, G. Q.; Eggan, K.; Hochedlinger, K.; Thomson, J.; Wang, W.; Gao, Y.; Zhang, K., Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **2009,** *27* (4), 353-60.
167. Huang, Y.; Pastor, W. A.; Shen, Y.; Tahiliani, M.; Liu, D. R.; Rao, A., The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS. One* **2010,** *5* (1), e8888.
168. Taub, F. E.; DeLeo, J. M.; Thompson, E. B., Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. *DNA* **1983,** *2* (4), 309-27.
169. Gitan, R. S.; Shi, H.; Chen, C. M.; Yan, P. S.; Huang, T. H., Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res* **2002,** *12* (1), 158-64.
170. Yan, P. S.; Wei, S. H.; Huang, T. H., Methylation-specific oligonucleotide microarray. *Methods Mol Biol* **2004,** *287*, 251-60.
171. Dedeurwaerder, S.; Defrance, M.; Calonne, E.; Denis, H.; Sotiriou, C.; Fuks, F., Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **2011,** *3* (6), 771-84.
172. Pidsley, R.; Zotenko, E.; Peters, T. J.; Lawrence, M. G.; Risbridger, G. P.; Molloy, P.; Van Djik, S.; Muhlhausler, B.; Stirzaker, C.; Clark, S. J., Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **2016,** *17* (1), 208.
173. Chen, Y. A.; Lemire, M.; Choufani, S.; Butcher, D. T.; Grafodatskaya, D.; Zanke, B. W.; Gallinger, S.; Hudson, T. J.; Weksberg, R., Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **2013,** *8* (2), 203-9.
174. Lizio, M.; Harshbarger, J.; Shimoji, H.; Severin, J.; Kasukawa, T.; Sahin, S.; Abugessaisa, I.; Fukuda, S.; Hori, F.; Ishikawa-Kato, S.; Mungall, C. J.; Arner, E.; Baillie, J. K.; Bertin, N.; Bono, H.; de Hoon, M.; Diehl, A. D.; Dimont, E.; Freeman, T. C.; Fujieda, K.; Hide, W.; Kaliyaperumal, R.; Katayama, T.; Lassmann, T.; Meehan, T. F.; Nishikata, K.; Ono, H.; Rehli, M.; Sandelin, A.; Schultes, E. A.; t Hoen, P. A.; Tatum, Z.; Thompson, M.; Toyoda, T.; Wright, D. W.; Daub, C. O.; Itoh, M.; Carninci, P.; Hayashizaki, Y.; Forrest, A. R.; Kawaji, H.; consortium, F., Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **2015,** *16*, 22.

175. Siggens, L.; Ekwall, K., Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. *J Intern Med* **2014,** *276* (3), 201-14.
176. Wong, N. C.; Ng, J.; Hall, N. E.; Lunke, S.; Salmanidis, M.; Brumatti, G.; Ekert, P. G.; Craig, J. M.; Saffery, R., Exploring the utility of human DNA methylation arrays for profiling mouse genomic DNA. *Genomics* **2013,** *102* (1), 38-46.
177. Needhamsen, M.; Ewing, E.; Lund, H.; Gomez-Cabrero, D.; Harris, R. A.; Kular, L.; Jagodic, M., Usability of human Infinium MethylationEPIC BeadChip for mouse DNA methylation studies. *BMC Bioinformatics* **2017,** *18* (1), 486.
178. Nazor, K. L.; Boland, M. J.; Bibikova, M.; Klotzle, B.; Yu, M.; Glenn-Pratola, V. L.; Schell, J. P.; Coleman, R. L.; Cabral-da-Silva, M. C.; Schmidt, U.; Peterson, S. E.; He, C.; Loring, J. F.; Fan, J. B., Application of a low cost array-based technique - TAB-Array - for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. *Genomics* **2014,** *104* (5), 358-67.
179. Canard, B.; Sarfati, R. S., DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* **1994,** *148* (1), 1-6.
180. Weber, M.; Davies, J. J.; Wittig, D.; Oakeley, E. J.; Haase, M.; Lam, W. L.; Schubeler, D., Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **2005,** *37* (8), 853-62.
181. Down, T. A.; Rakyan, V. K.; Turner, D. J.; Flicek, P.; Li, H.; Kulesha, E.; Graf, S.; Johnson, N.; Herrero, J.; Tomazou, E. M.; Thorne, N. P.; Backdahl, L.; Herberth, M.; Howe, K. L.; Jackson, D. K.; Miretti, M. M.; Marioni, J. C.; Birney, E.; Hubbard, T. J.; Durbin, R.; Tavare, S.; Beck, S., A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **2008,** *26* (7), 779-85.
182. Shen, L.; Wu, H.; Diep, D.; Yamaguchi, S.; D'Alessio, A. C.; Fung, H. L.; Zhang, K.; Zhang, Y., Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **2013,** *153* (3), 692-706.
183. Pastor, W. A.; Pape, U. J.; Huang, Y.; Henderson, H. R.; Lister, R.; Ko, M.; McLoughlin, E. M.; Brudno, Y.; Mahapatra, S.; Kapranov, P.; Tahiliani, M.; Daley, G. Q.; Liu, X. S.; Ecker, J. R.; Milos, P. M.; Agarwal, S.; Rao, A., Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **2011,** *473* (7347), 394-7.
184. Huang, Y.; Pastor, W. A.; Zepeda-Martinez, J. A.; Rao, A., The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat Protoc* **2012,** *7* (10), 1897-908.
185. Song, C. X.; Sun, Y.; Dai, Q.; Lu, X. Y.; Yu, M.; Yang, C. G.; He, C., Detection of 5-hydroxymethylcytosine in DNA by transferring a keto-glucose by using T4 phage beta-glucosyltransferase. *Chembiochem* **2011,** *12* (11), 1682-1685.
186. Song, C. X.; Szulwach, K. E.; Fu, Y.; Dai, Q.; Yi, C.; Li, X.; Li, Y.; Chen, C. H.; Zhang, W.; Jian, X.; Wang, J.; Zhang, L.; Looney, T. J.; Zhang, B.; Godley, L. A.; Hicks, L. M.; Lahn, B. T.; Jin, P.; He, C., Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol* **2011,** *29* (1), 68-72.

187. Song, C. X.; Szulwach, K. E.; Dai, Q.; Fu, Y.; Mao, S. Q.; Lin, L.; Street, C.; Li, Y.; Poidevin, M.; Wu, H.; Gao, J.; Liu, P.; Li, L.; Xu, G. L.; Jin, P.; He, C., Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **2013,** *153* (3), 678-691.
188. Yong, W. S.; Hsu, F. M.; Chen, P. Y., Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **2016,** *9*, 26.
189. Gupta, R.; Nagarajan, A.; Wajapeyee, N., Advances in genome-wide DNA methylation analysis. *Biotechniques* **2010,** *49* (4), iii-xi.
190. Frommer, M.; McDonald, L. E.; Millar, D. S.; Collis, C. M.; Watt, F.; Grigg, G. W.; Molloy, P. L.; Paul, C. L., A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **1992,** *89* (5), 1827-31.
191. Stirzaker, C.; Taberlay, P. C.; Statham, A. L.; Clark, S. J., Mining cancer methylomes: prospects and challenges. *Trends Genet* **2014,** *30* (2), 75-84.
192. Lister, R.; O'Malley, R. C.; Tonti-Filippini, J.; Gregory, B. D.; Berry, C. C.; Millar, A. H.; Ecker, J. R., Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **2008,** *133* (3), 523-36.
193. Cokus, S. J.; Feng, S.; Zhang, X.; Chen, Z.; Merriman, B.; Haudenschild, C. D.; Pradhan, S.; Nelson, S. F.; Pellegrini, M.; Jacobsen, S. E., Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **2008,** *452* (7184), 215-9.
194. Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q. M.; Edsall, L.; Antosiewicz-Bourget, J.; Stewart, R.; Ruotti, V.; Millar, A. H.; Thomson, J. A.; Ren, B.; Ecker, J. R., Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **2009,** *462* (7271), 315-22.
195. Song, Q.; Decato, B.; Hong, E. E.; Zhou, M.; Fang, F.; Qu, J.; Garvin, T.; Kessler, M.; Zhou, J.; Smith, A. D., A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **2013,** *8* (12), e81148.
196. Wang, L.; Sun, J.; Wu, H.; Liu, S.; Wang, J.; Wu, B.; Huang, S.; Li, N.; Wang, J.; Zhang, X., Systematic assessment of reduced representation bisulfite sequencing to human blood samples: A promising method for large-sample-scale epigenomic studies. *J Biotechnol* **2012,** *157* (1), 1-6.
197. Meissner, A.; Mikkelsen, T. S.; Gu, H.; Wernig, M.; Hanna, J.; Sivachenko, A.; Zhang, X.; Bernstein, B. E.; Nusbaum, C.; Jaffe, D. B.; Gnirke, A.; Jaenisch, R.; Lander, E. S., Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **2008,** *454* (7205), 766-70.
198. Gu, H.; Bock, C.; Mikkelsen, T. S.; Jager, N.; Smith, Z. D.; Tomazou, E.; Gnirke, A.; Lander, E. S.; Meissner, A., Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods* **2010,** *7* (2), 133-6.

199. Gu, H.; Smith, Z. D.; Bock, C.; Boyle, P.; Gnirke, A.; Meissner, A., Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **2011,** *6* (4), 468-81.
200. Doi, A.; Park, I. H.; Wen, B.; Murakami, P.; Aryee, M. J.; Irizarry, R.; Herb, B.; Ladd-Acosta, C.; Rho, J.; Loewer, S.; Miller, J.; Schlaeger, T.; Daley, G. Q.; Feinberg, A. P., Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* **2009,** *41* (12), 1350-3.
201. Wang, J.; Xia, Y.; Li, L.; Gong, D.; Yao, Y.; Luo, H.; Lu, H.; Yi, N.; Wu, H.; Zhang, X.; Tao, Q.; Gao, F., Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. *BMC Genomics* **2013,** *14*, 11.
202. Lee, Y. K.; Jin, S.; Duan, S.; Lim, Y. C.; Ng, D. P.; Lin, X. M.; Yeo, G. S.; Ding, C., Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples. *Biol Proced Online* **2014,** *16* (1), 1.
203. Chen, K.; Zhao, B. S.; He, C., Nucleic Acid Modifications in Regulation of Gene Expression. *Cell Chem Biol* **2016,** *23* (1), 74-85.
204. Jin, S. G.; Kadam, S.; Pfeifer, G. P., Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **2010,** *38* (11), e125.
205. Booth, M. J.; Branco, M. R.; Ficz, G.; Oxley, D.; Krueger, F.; Reik, W.; Balasubramanian, S., Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **2012,** *336* (6083), 934-7.
206. Yu, M.; Hon, G. C.; Szulwach, K. E.; Song, C. X.; Zhang, L.; Kim, A.; Li, X.; Dai, Q.; Shen, Y.; Park, B.; Min, J. H.; Jin, P.; Ren, B.; He, C., Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **2012,** *149* (6), 1368-1380.
207. Yu, M.; Hon, G. C.; Szulwach, K. E.; Song, C. X.; Jin, P.; Ren, B.; He, C., Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protoc* **2012,** *7* (12), 2159-2170.
208. Chen, K.; Zhang, J.; Guo, Z.; Ma, Q.; Xu, Z.; Zhou, Y.; Xu, Z.; Li, Z.; Liu, Y.; Ye, X.; Li, X.; Yuan, B.; Ke, Y.; He, C.; Zhou, L.; Liu, J.; Ci, W., Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res* **2016,** *26* (1), 103-18.
209. Booth, M. J.; Marsico, G.; Bachman, M.; Beraldi, D.; Balasubramanian, S., Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **2014,** *6* (5), 435-40.
210. Wu, H.; Wu, X.; Zhang, Y., Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. *Nat Protoc* **2016,** *11* (6), 1081-100.
211. Fu, Y.; Luo, G. Z.; Chen, K.; Deng, X.; Yu, M.; Han, D.; Hao, Z.; Liu, J.; Lu, X.; Dore, L. C.; Weng, X.; Ji, Q.; Mets, L.; He, C., N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **2015,** *161* (4), 879-892.

212. Nakano, K.; Shiroma, A.; Shimoji, M.; Tamotsu, H.; Ashimine, N.; Ohki, S.; Shinzato, M.; Minami, M.; Nakanishi, T.; Teruya, K.; Satou, K.; Hirano, T., Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum Cell* **2017,** *30* (3), 149-161.
213. Rhoads, A.; Au, K. F., PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **2015,** *13* (5), 278-89.
214. Clark, T. A.; Lu, X.; Luong, K.; Dai, Q.; Boitano, M.; Turner, S. W.; He, C.; Korlach, J., Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol* **2013,** *11*, 4.
215. Chavez, L.; Huang, Y.; Luong, K.; Agarwal, S.; Iyer, L. M.; Pastor, W. A.; Hench, V. K.; Frazier-Bowers, S. A.; Korol, E.; Liu, S.; Tahiliani, M.; Wang, Y.; Clark, T. A.; Korlach, J.; Pukkila, P. J.; Aravind, L.; Rao, A., Simultaneous sequencing of oxidized methylcytosines produced by TET/JBP dioxygenases in Coprinopsis cinerea. *Proc Natl Acad Sci U S A* **2014,** *111* (48), E5149-58.
216. Niedringhaus, T. P.; Milanova, D.; Kerby, M. B.; Snyder, M. P.; Barron, A. E., Landscape of next-generation sequencing technologies. *Anal Chem* **2011,** *83* (12), 4327-41.
217. Wescoe, Z. L.; Schreiber, J.; Akeson, M., Nanopores discriminate among five C5-cytosine variants in DNA. *J Am Chem Soc* **2014,** *136* (47), 16582-7.
218. Rand, A. C.; Jain, M.; Eizenga, J. M.; Musselman-Brown, A.; Olsen, H. E.; Akeson, M.; Paten, B., Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* **2017,** *14* (4), 411-413.
219. Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (accessed March 03, 2018).
220. Ziller, M. J.; Hansen, K. D.; Meissner, A.; Aryee, M. J., Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* **2015,** *12* (3), 230-2, 1 p following 232.
221. Hing, B.; Ramos, E.; Braun, P.; McKane, M.; Jancic, D.; Tamashiro, K. L.; Lee, R. S.; Michaelson, J. J.; Druley, T. E.; Potash, J. B., Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tissue-specific DNA methylation patterns of genes involved in neurodevelopment. *Epigenetics* **2015,** *10* (7), 581-96.
222. Fu, Y.; Springer, N. M.; Ying, K.; Yeh, C. T.; Iniguez, A. L.; Richmond, T.; Wu, W.; Barbazuk, B.; Nettleton, D.; Jeddeloh, J.; Schnable, P. S., High-resolution genotyping via whole genome hybridizations to microarrays containing long oligonucleotide probes. *PLoS One* **2010,** *5* (12), e14178.
223. Li, Q.; Suzuki, M.; Wendt, J.; Patterson, N.; Eichten, S. R.; Hermanson, P. J.; Green, D.; Jeddeloh, J.; Richmond, T.; Rosenbaum, H.; Burgess, D.; Springer, N. M.; Greally, J. M., Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic Acids Res* **2015**.
224. Qian, X.; Melkamu, T.; Upadhyaya, P.; Kassie, F., Indole-3-carbinol inhibited tobacco smoke carcinogen-induced lung adenocarcinoma in A/J mice when administered

during the post-initiation or progression phase of lung tumorigenesis. *Cancer Lett* **2011,** *311* (1), 57-65.
225. Lin, C.; Song, H.; Huang, C.; Yao, E.; Gacayan, R.; Xu, S. M.; Chuang, P. T., Alveolar type II cells possess the capability of initiating lung tumor development. *PLoS One* **2012,** *7* (12), e53817.
226. Hendrich, B.; Bird, A., Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* **1998,** *18* (11), 6538-47.
227. Nan, X.; Ng, H. H.; Johnson, C. A.; Laherty, C. D.; Turner, B. M.; Eisenman, R. N.; Bird, A., Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **1998,** *393* (6683), 386-389.
228. Baylin, S. B.; Jones, P. A., A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* **2011,** *11* (10), 726-34.
229. Goll, M. G.; Bestor, T. H., Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **2005,** *74*, 481-514.
230. Jacinto, F. V.; Esteller, M., Mutator pathways unleashed by epigenetic silencing in human cancer. *Mutagenesis* **2007,** *22* (4), 247-53.
231. Valinluck, V.; Sowers, L. C., Inflammation-mediated cytosine damage: a mechanistic link between inflammation and the epigenetic alterations in human cancers. *Cancer Res* **2007,** *67* (12), 5583-5586.
232. Valinluck, V.; Liu, P.; Kang, J. I., Jr.; Burdzy, A.; Sowers, L. C., 5-halogenated pyrimidine lesions within a CpG sequence context mimic 5-methylcytosine by enhancing the binding of the methyl-CpG-binding domain of methyl-CpG-binding protein 2 (MeCP2). *Nucleic Acids Res* **2005,** *33* (9), 3057-3064.
233. Ji, D.; Lin, K.; Song, J.; Wang, Y., Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. *Mol Biosyst* **2014,** *10* (7), 1749-52.
234. Gait, M. J., *Oligonucleotide synthesis: a practical approach*. IRL Press: Washington, DC, 1984.
235. Matter, B.; Wang, G.; Jones, R.; Tretyakova, N., Formation of diastereomeric benzo[*a*]pyrene diol epoxide-guanine adducts in *p53* gene-derived DNA sequences. *Chem. Res. Toxicol* **2004,** *17* (6), 731-741.
236. Ziegel, R.; Shallop, A.; Upadhyaya, P.; Jones, R.; Tretyakova, N., Endogenous 5-methylcytosine protects neighboring guanines from N7 and O^6-methylation and *O6*-pyridyloxobutylation by the tobacco carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Biochemistry* **2004,** *43* (2), 540-549.
237. Tretyakova, N.; Matter, B.; Jones, R.; Shallop, A., Formation of benzo[*a*]pyrene diol epoxide-DNA adducts at specific guanines within *K-ras* and *p53* gene sequences: stable isotope-labeling mass spectrometry approach. *Biochemistry* **2002,** *41* (30), 9535-9544.
238. Ziegel, R.; Shallop, A.; Jones, R.; Tretyakova, N., *K-ras* gene sequence effects on the formation of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)-DNA adducts. *Chem. Res. Toxicol* **2003,** *16* (4), 541-550.

239. Guza, R.; Rajesh, M.; Fang, Q.; Pegg, A. E.; Tretyakova, N., Kinetics of O^6-Me-dG repair by O^6-alkylguanine DNA-alkyltransferase within *K-ras* gene derived DNA sequences. *Chem. Res. Toxicol* **2006,** *19*, 531-538.
240. Guza, R.; Ma, L.; Fang, Q.; Pegg, A. E.; Tretyakova, N. Y., Cytosine methylation effects on the repair of O^6-methylguanines within CG dinucleotides. *J. Biol. Chem* **2009,** *284* (34), 22601-22610.
241. Rajesh, M.; Wang, G.; Jones, R.; Tretyakova, N., Stable isotope labeling-mass spectrometry analysis of methyl- and pyridyloxobutyl-guanine adducts of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone in *p53*-derived DNA sequences. *Biochemistry* **2005,** *44* (6), 2197-2207.
242. Guza, R.; Kotandeniya, D.; Murphy, K.; Dissanayake, T.; Lin, C.; Giambasu, G. M.; Lad, R. R.; Wojciechowski, F.; Amin, S.; Sturla, S. J.; Hudson, R. H.; York, D. M.; Jankowiak, R.; Jones, R.; Tretyakova, N. Y., Influence of C-5 substituted cytosine and related nucleoside analogs on the formation of benzo[*a*]pyrene diol epoxide-dG adducts at CG base pairs of DNA. *Nucleic Acids Res* **2011,** *39* (9), 3988-4006.
243. *Schrodinger Modeling Suite Package*, 10.3; Schrodinger LLC, New York, NY: 2015.
244. Song, J.; Rechkoblit, O.; Bestor, T. H.; Patel, D. J., Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **2011,** *331* (6020), 1036-40.
245. Wickramaratne, S.; Ji, S.; Mukherjee, S.; Su, Y.; Pence, M. G.; Lior-Hoffmann, L.; Fu, I.; Broyde, S.; Guengerich, F. P.; Distefano, M.; Scharer, O. D.; Sham, Y. Y.; Tretyakova, N., Bypass of DNA-Protein Cross-links Conjugated to the 7-Deazaguanine Position of DNA by Translesion Synthesis Polymerases. *J Biol Chem* **2016,** *291* (45), 23589-23603.
246. Tang, J.; Maddali, K.; Pommier, Y.; Sham, Y. Y.; Wang, Z., Scaffold rearrangement of dihydroxypyrimidine inhibitors of HIV integrase: Docking model revisited. *Bioorg Med Chem Lett* **2010,** *20* (11), 3275-9.
247. Vyas, V. K.; Ukawala, R. D.; Ghate, M.; Chintha, C., Homology modeling a fast tool for drug discovery: current perspectives. *Indian J Pharm Sci* **2012,** *74* (1), 1-17.
248. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **1996,** *118* (45), 11225-11236.
249. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **1990,** *112* (16), 6127-6129.
250. Robins, M. J.; Barr, P. J.; Giziewicz, J., Nucleic acid related compounds. 38. Smooth and high-yield iodination and chlorination at C-5 of uracil bases and p-toluyl-protected nuclleosides. *Can. J. Chem* **1982,** *60* (5), 554-557.
251. Robins, M. J.; Barr, P. J., Nucleic acid related compounds. 39. Efficient conversion of 5-iodo to 5-alkynyl and derived 5-substituted uracil bases and nucleosides. *J. Org. Chem* **1983,** *48*, 1854-1862.

252. Sonogashira, K., Development of Pd–Cu catalyzed cross-coupling of terminal acetylenes with sp2-carbon halides. *Journal of Organometallic Chemistry* **2002,** *653* (1-2), 46-49.
253. Sonogashira, K.; Tohda, Y.; Hagihara, N., A convenient synthesis of acetylenes: catalytic substitutions of acetylenic hydrogen with bromoalkenes, iodoarenes and bromopyridines. *Tetrahed. Lett* **1975,** *16* (50), 4467-4470.
254. Denissenko, M. F.; Chen, J. X.; Tang, M. S.; Pfeifer, G. P., Cytosine methylation determines hot spots of DNA damage in the human *p53* gene. *Proc. Natl. Acad. Sci. U. S. A* **1997,** *94* (8), 3893-3898.
255. Gromova, E. S.; Subach, O. M.; Baskunov, V. B.; Geacintov, N., Impact of Carcinogen-DNA Adducts on DNA Methylation. In *Structural Biology of DNA Damage and Repair*, Stone, M. P., Ed. American Chemical Society: Washington, DC, 2010; pp 103-116.
256. Yang, J.; Lior-Hoffmann, L.; Wang, S.; Zhang, Y.; Broyde, S., DNA cytosine methylation: structural and thermodynamic characterization of the epigenetic marking mechanism. *Biochemistry* **2013,** *52* (16), 2828-38.
257. Jin, B.; Li, Y.; Robertson, K. D., DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2011,** *2* (6), 607-617.
258. Jones, P. A.; Takai, D., The role of DNA methylation in mammalian epigenetics. *Science* **2001,** *293* (5532), 1068-1070.
259. Sharif, J.; Koseki, H., Recruitment of Dnmt1 roles of the SRA protein Np95 (Uhrf1) and other factors. *Prog Mol Biol Transl Sci* **2011,** *101*, 289-310.
260. Matje, D. M.; Zhou, H.; Smith, D. A.; Neely, R. K.; Dryden, D. T.; Jones, A. C.; Dahlquist, F. W.; Reich, N. O., Enzyme-promoted base flipping controls DNA methylation fidelity. *Biochemistry* **2013,** *52* (10), 1677-85.
261. Matje, D. M.; Krivacic, C. T.; Dahlquist, F. W.; Reich, N. O., Distal structural elements coordinate a conserved base flipping network. *Biochemistry* **2013,** *52* (10), 1669-76.
262. Svedruzic, Z. M.; Reich, N. O., The mechanism of target base attack in DNA cytosine carbon 5 methylation. *Biochemistry* **2004,** *43* (36), 11460-73.
263. Maiti, A.; Drohat, A. C., Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem* **2011,** *286* (41), 35334-35338.
264. Seisenberger, S.; Peat, J. R.; Reik, W., Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. *Curr Opin Cell Biol* **2013,** *25* (3), 281-8.
265. Wang, L.; Zhang, J.; Duan, J.; Gao, X.; Zhu, W.; Lu, X.; Yang, L.; Zhang, J.; Li, G.; Ci, W.; Li, W.; Zhou, Q.; Aluru, N.; Tang, F.; He, C.; Huang, X.; Liu, J., Programming and inheritance of parental DNA methylomes in mammals. *Cell* **2014,** *157* (4), 979-991.
266. Pradhan, M.; Esteve, P. O.; Chin, H. G.; Samaranayke, M.; Kim, G. D.; Pradhan, S., CXXC domain of human DNMT1 is essential for enzymatic activity. *Biochemistry* **2008,** *47* (38), 10000-10009.

267. Chuang, L. S.; Ian, H. I.; Koh, T. W.; Ng, H. H.; Xu, G.; Li, B. F., Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science* **1997,** *277* (5334), 1996-2000.
268. Kim, G. D.; Ni, J.; Kelesoglu, N.; Roberts, R. J.; Pradhan, S., Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferases. *EMBO J* **2002,** *21* (15), 4183-95.
269. D. E. Shaw Research, N. Y., NY *Desmond Molecular Dynamics System*, 2018.1; 2018.
270. W.L., J. J. C. J. D. M. R. W. I. M. L. K., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79*, 926.
271. Zhou, Y. B.; Wu, Y. Y.; Pokholenko, O.; Grimsrud, M.; Sham, Y.; Papper, V.; Marks, R.; Steele, T., Aptamer adaptive binding assessed by stilbene photoisomerization towards regenerating aptasensors. *Sensor Actuat B-Chem* **2018,** *257*, 245-255.
272. Pradhan, S.; Bacolla, A.; Wells, R. D.; Roberts, R. J., Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J Biol Chem* **1999,** *274* (46), 33002-10.
273. Kohli, R. M.; Zhang, Y., TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **2013,** *502* (7472), 472-479.
274. Mayer, W.; Niveleau, A.; Walter, J.; Fundele, R.; Haaf, T., Demethylation of the zygotic paternal genome. *Nature* **2000,** *403* (6769), 501-2.
275. Lister, R.; Mukamel, E. A., Turning over DNA methylation in the mind. *Front Neurosci* **2015,** *9*, 252.
276. Lister, R.; Mukamel, E. A.; Nery, J. R.; Urich, M.; Puddifoot, C. A.; Johnson, N. D.; Lucero, J.; Huang, Y.; Dwork, A. J.; Schultz, M. D.; Yu, M.; Tonti-Filippini, J.; Heyn, H.; Hu, S.; Wu, J. C.; Rao, A.; Esteller, M.; He, C.; Haghighi, F. G.; Sejnowski, T. J.; Behrens, M. M.; Ecker, J. R., Global epigenomic reconfiguration during mammalian brain development. *Science* **2013,** *341* (6146), 1237905.
277. Aranda, J.; Zinovjev, K.; Swiderek, K.; Roca, M.; Tunon, I., Unraveling the Reaction Mechanism of Enzymatic C5-Cytosine Methylation of DNA. A Combined Molecular Dynamics and QM/MM Study of Wild Type and Gln119 Variant. *Acs Catal* **2016,** *6* (5), 3262-3276.
278. Deaton, A. M.; Bird, A., CpG islands and the regulation of transcription. *Genes Dev* **2011,** *25* (10), 1010-1022.
279. Jeziorska, D. M.; Murray, R. J. S.; De Gobbi, M.; Gaentzsch, R.; Garrick, D.; Ayyub, H.; Chen, T.; Li, E.; Telenius, J.; Lynch, M.; Graham, B.; Smith, A. J. H.; Lund, J. N.; Hughes, J. R.; Higgs, D. R.; Tufarelli, C., DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *Proc Natl Acad Sci U S A* **2017,** *114* (36), E7526-E7535.
280. Robertson, K. D., DNA methylation and human disease. *Nat. Rev. Genet* **2005,** *6* (8), 597-610.
281. Baylin, S. B.; Jones, P. A., Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **2016,** *8* (9).

282. Adcock, I. M.; Tsaprouni, L.; Bhavsar, P.; Ito, K., Epigenetic regulation of airway inflammation. *Curr Opin Immunol* **2007,** *19* (6), 694-700.
283. Korsgren, M.; Linden, M.; Entwistle, N.; Cook, J.; Wollmer, P.; Andersson, M.; Larsson, B.; Greiff, L., Inhalation of LPS induces inflammatory airway responses mimicking characteristics of chronic obstructive pulmonary disease. *Clin. Physiol Funct. Imaging* **2012,** *32* (1), 71-79.
284. Bowman, R. V.; Wright, C. M.; Davidson, M. R.; Francis, S. M.; Yang, I. A.; Fong, K. M., Epigenomic targets for the treatment of respiratory disease. *Expert Opin Ther Targets* **2009,** *13* (6), 625-40.
285. Bosetti, C.; Gallus, S.; La Vecchia, C., Aspirin and cancer risk: a summary review to 2007. *Recent Results Cancer Res* **2009,** *181*, 231-51.
286. Oh, S. W.; Myung, S. K.; Park, J. Y.; Lee, C. M.; Kwon, H. T.; Korean Meta-analysis Study, G., Aspirin use and risk for lung cancer: a meta-analysis. *Ann Oncol* **2011,** *22* (11), 2456-65.
287. McCormack, V. A.; Hung, R. J.; Brenner, D. R.; Bickeboller, H.; Rosenberger, A.; Muscat, J. E.; Lazarus, P.; Tjonneland, A.; Friis, S.; Christiani, D. C.; Chun, E. M.; Le Marchand, L.; Rennert, G.; Rennert, H. S.; Andrew, A. S.; Orlow, I.; Park, B.; Boffetta, P.; Duell, E. J., Aspirin and NSAID use and lung cancer risk: a pooled analysis in the International Lung Cancer Consortium (ILCCO). *Cancer Causes Control* **2011,** *22* (12), 1709-20.
288. Xu, J.; Yin, Z.; Gao, W.; Liu, L.; Wang, R.; Huang, P.; Yin, Y.; Liu, P.; Yu, R.; Shu, Y., Meta-analysis on the association between nonsteroidal anti-inflammatory drug use and lung cancer risk. *Clin Lung Cancer* **2012,** *13* (1), 44-51.
289. Shechter, D.; Dormann, H. L.; Allis, C. D.; Hake, S. B., Extraction, purification and analysis of histones. *Nat Protoc* **2007,** *2* (6), 1445-57.
290. Zhou, T.; Chung, Y. H.; Chen, J.; Chen, Y., Site-Specific Identification of Lysine Acetylation Stoichiometries in Mammalian Cells. *J Proteome Res* **2016,** *15* (3), 1103-13.
291. Izzotti, A.; Balansky, R.; Ganchev, G.; Iltcheva, M.; Longobardi, M.; Pulliero, A.; Camoirano, A.; D'Agostini, F.; Geretto, M.; Micale, R. T.; La Maestra, S.; Miller, M. S.; Steele, V. E.; De Flora, S., Early and late effects of aspirin and naproxen on microRNAs in the lung and blood of mice, either unexposed or exposed to cigarette smoke. *Oncotarget* **2017,** *8* (49), 85716-85748.
292. Qu, J.; Zhou, M.; Song, Q.; Hong, E. E.; Smith, A. D., MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **2013,** *29* (20), 2645-6.
293. Dolzhenko, E.; Smith, A. D., Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **2014,** *15*, 215.
294. Edwards, S. H.; Rossiter, L. M.; Taylor, K. M.; Holman, M. R.; Zhang, L.; Ding, Y. S.; Watson, C. H., Tobacco-Specific Nitrosamines in the Tobacco and Mainstream Smoke of U.S. Commercial Cigarettes. *Chem Res Toxicol* **2017,** *30* (2), 540-551.

295. Lu, Y. C.; Yeh, W. C.; Ohashi, P. S., LPS/TLR4 signal transduction pathway. *Cytokine* **2008,** *42* (2), 145-151.
296. Bojesen, S. E.; Timpson, N.; Relton, C.; Davey Smith, G.; Nordestgaard, B. G., AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax* **2017,** *72* (7), 646-653.
297. Strahl, B. D.; Allis, C. D., The language of covalent histone modifications. *Nature* **2000,** *403* (6765), 41-5.
298. Fraga, M. F.; Ballestar, E.; Villar-Garea, A.; Boix-Chornet, M.; Espada, J.; Schotta, G.; Bonaldi, T.; Haydon, C.; Ropero, S.; Petrie, K.; Iyer, N. G.; Perez-Rosado, A.; Calvo, E.; Lopez, J. A.; Cano, A.; Calasanz, M. J.; Colomer, D.; Piris, M. A.; Ahn, N.; Imhof, A.; Caldas, C.; Jenuwein, T.; Esteller, M., Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nat Genet* **2005,** *37* (4), 391-400.
299. van Leeuwen, F.; van Steensel, B., Histone modifications: from genome-wide maps to functional insights. *Genome Biol* **2005,** *6* (6), 113.
300. Lee, K. K.; Workman, J. L., Histone acetyltransferase complexes: one size doesn't fit all. *Nat Rev Mol Cell Biol* **2007,** *8* (4), 284-95.
301. Karmodiya, K.; Krebs, A. R.; Oulad-Abdelghani, M.; Kimura, H.; Tora, L., H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **2012,** *13*, 424.
302. Bodai, L.; Zsindely, N.; Gaspar, R.; Kristo, I.; Komonyi, O.; Boros, I. M., Ecdysone induced gene expression is associated with acetylation of histone H3 lysine 23 in Drosophila melanogaster. *PLoS One* **2012,** *7* (7), e40565.
303. Cai, L.; Sutter, B. M.; Li, B.; Tu, B. P., Acetyl-CoA induces cell growth and proliferation by promoting the acetylation of histones at growth genes. *Mol Cell* **2011,** *42* (4), 426-37.
304. Kim, Y. Z., Altered histone modifications in gliomas. *Brain Tumor Res Treat* **2014,** *2* (1), 7-21.
305. Society, A. C., Cancer Facts and Figures 2015. 2015.
306. Prevention, C. f. D. C. a., Current Cigarette Smoking Among Adults. 2015; pp 1233-1240.
307. Hecht, S. S., Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat. Rev. Cancer* **2003,** *3* (10), 733-744.
308. Peterson, L. A., Formation, repair, and genotoxic properties of bulky DNA adducts formed from tobacco-specific nitrosamines. *J Nucleic Acids* **2010,** *2010*.
309. Shiels, M. S.; Pfeiffer, R. M.; Hildesheim, A.; Engels, E. A.; Kemp, T. J.; Park, J. H.; Katki, H. A.; Koshiol, J.; Shelton, G.; Caporaso, N. E.; Pinto, L. A.; Chaturvedi, A. K., Circulating inflammation markers and prospective risk for lung cancer. *J. Natl. Cancer Inst* **2013,** *105* (24), 1871-1880.
310. Maekita, T.; Nakazawa, K.; Mihara, M.; Nakajima, T.; Yanaoka, K.; Iguchi, M.; Arii, K.; Kaneda, A.; Tsukamoto, T.; Tatematsu, M.; Tamura, G.; Saito, D.; Sugimura, T.;

Ichinose, M.; Ushijima, T., High levels of aberrant DNA methylation in Helicobacter pylori-infected gastric mucosae and its possible association with gastric cancer risk. *Clin. Cancer Res* **2006,** *12* (3 Pt 1), 989-995.
311. Chiba, T.; Marusawa, H.; Ushijima, T., Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. *Gastroenterology* **2012,** *143* (3), 550-563.
312. Nishida, N.; Kudo, M., Oxidative stress and epigenetic instability in human hepatocarcinogenesis. *Dig. Dis* **2013,** *31* (5-6), 447-453.
313. Cao, Y.; Nishihara, R.; Wu, K.; Wang, M.; Ogino, S.; Willett, W. C.; Spiegelman, D.; Fuchs, C. S.; Giovannucci, E. L.; Chan, A. T., Population-wide Impact of Long-term Use of Aspirin and the Risk for Cancer. *JAMA Oncol* **2016,** *2* (6), 762-9.
314. Jalbert, G.; Castonguay, A., Effects of NSAIDs on NNK-induced pulmonary and gastric tumorigenesis in A/J mice. *Cancer Lett* **1992,** *66* (1), 21-8.
315. Belinsky, S. A.; Swafford, D. S.; Finch, G. L.; Mitchell, C. E.; Kelly, G.; Hahn, F. F.; Anderson, M. W.; Nikula, K. J., Alterations in the *K-ras* and *p53* genes in rat lung tumors. *Environ. Health Perspect* **1997,** *105 Suppl 4*, 901-906.
316. Floyd, H. S.; Jennings-Gee, J. E.; Kock, N. D.; Miller, M. S., Genetic and epigenetic alterations in lung tumors from bitransgenic *Ki-ras* G12C expressing mice. *Mol. Carcinog* **2006,** *45* (7), 506-517.
317. Hollstein, M.; Rice, K.; Greenblatt, M. S.; Soussi, T.; Fuchs, R.; Sorlie, T.; Hovig, E.; Smith-Sorensen, B.; Montesano, R.; Harris, C. C., Database of *p53* gene somatic mutations in human tumors and cell lines. *Nucleic Acids Res* **1994,** *22* (17), 3551-3555.
318. Ushijima, T.; Asada, K., Aberrant DNA methylation in contrast with mutations. *Cancer Sci* **2010,** *101* (2), 300-305.
319. Ballestar, E.; Esteller, M., The impact of chromatin in human cancer: linking DNA methylation to gene silencing. *Carcinogenesis* **2002,** *23* (7), 1103-1109.
320. Cho, B.; Lee, H.; Jeong, S.; Bang, Y. J.; Lee, H. J.; Hwang, K. S.; Kim, H. Y.; Lee, Y. S.; Kang, G. H.; Jeoung, D. I., Promoter hypomethylation of a novel cancer/testis antigen gene CAGE is correlated with its aberrant expression and is seen in premalignant stage of gastric carcinoma. *Biochem. Biophys. Res. Commun* **2003,** *307* (1), 52-63.
321. Esteller, M.; Hamilton, S. R.; Burger, P. C.; Baylin, S. B.; Herman, J. G., Inactivation of the DNA repair gene O^6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Res* **1999,** *59* (4), 793-797.
322. Fujiwara, K.; Fujimoto, N.; Tabata, M.; Nishii, K.; Matsuo, K.; Hotta, K.; Kozuki, T.; Aoe, M.; Kiura, K.; Ueoka, H.; Tanimoto, M., Identification of epigenetic aberrant promoter methylation in serum DNA is useful for early detection of lung cancer. *Clin. Cancer Res* **2005,** *11* (3), 1219-1225.
323. Liloglou, T.; Bediaga, N. G.; Brown, B. R.; Field, J. K.; Davies, M. P., Epigenetic biomarkers in lung cancer. *Cancer Lett* **2014,** *342* (2), 200-212.

324. Wielscher, M.; Liou, W.; Pulverer, W.; Singer, C. F.; Rappaport-Fuerhauser, C.; Kandioler, D.; Egger, G.; Weinhausel, A., Cytosine 5-Hydroxymethylation of the LZTS1 Gene Is Reduced in Breast Cancer. *Transl. Oncol* **2013,** *6* (6), 715-721.
325. Shen, H.; Laird, P. W., Interplay between the cancer genome and epigenome. *Cell* **2013,** *153* (1), 38-55.
326. Rawluszko-Wieczorek, A. A.; Siera, A.; Jagodzinski, P. P., TET proteins in cancer: Current 'state of the art'. *Crit Rev Oncol Hematol* **2015,** *96* (3), 425-36.
327. Carey, N.; Marques, C. J.; Reik, W., DNA demethylases: a new epigenetic frontier in drug discovery. *Drug Discov. Today* **2011,** *16* (15-16), 683-690.
328. Pulling, L. C.; Divine, K. K.; Klinge, D. M.; Gilliland, F. D.; Kang, T.; Schwartz, A. G.; Bocklage, T. J.; Belinsky, S. A., Promoter hypermethylation of the O^6-methylguanine-DNA methyltransferase gene: more common in lung adenocarcinomas from never-smokers than smokers and associated with tumor progression. *Cancer Res* **2003,** *63* (16), 4842-4848.
329. Liu, S., Epigenetics advancing personalized nanomedicine in cancer therapy. *Adv Drug Deliv Rev* **2012,** *64* (13), 1532-43.
330. Muller, U.; Bauer, C.; Siegl, M.; Rottach, A.; Leonhardt, H., TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation. *Nucleic Acids Res* **2014,** *42* (13), 8592-8604.
331. Jones, P. A.; Baylin, S. B., The epigenomics of cancer. *Cell* **2007,** *128* (4), 683-692.
332. Zhang, Q. H.; Dai, X. H.; Dai, Z. M.; Cai, Y. N., Genome-scale meta-analysis of DNA methylation during progression of lung adenocarcinoma. *Genet. Mol. Res* **2015,** *14* (3), 9200-9214.
333. Belinsky, S. A., Silencing of genes by promoter hypermethylation: key event in rodent and human lung cancer. *Carcinogenesis* **2005,** *26* (9), 1481-1487.
334. Yang, H.; Liu, Y.; Bai, F.; Zhang, J. Y.; Ma, S. H.; Liu, J.; Xu, Z. D.; Zhu, H. G.; Ling, Z. Q.; Ye, D.; Guan, K. L.; Xiong, Y., Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* **2013,** *32* (5), 663-669.
335. Brock, M. V.; Hooker, C. M.; Ota-Machida, E.; Han, Y.; Guo, M.; Ames, S.; Glockner, S.; Piantadosi, S.; Gabrielson, E.; Pridham, G.; Pelosky, K.; Belinsky, S. A.; Yang, S. C.; Baylin, S. B.; Herman, J. G., DNA methylation markers and early recurrence in stage I lung cancer. *N. Engl. J. Med* **2008,** *358* (11), 1118-1128.
336. Kontic, M.; Stojsic, J.; Jovanovic, D.; Bunjevacki, V.; Ognjanovic, S.; Kuriger, J.; Puumala, S.; Nelson, H. H., Aberrant promoter methylation of CDH13 and MGMT genes is associated with clinicopathologic characteristics of primary non-small-cell lung carcinoma. *Clin. Lung Cancer* **2012,** *13* (4), 297-303.
337. Lin, Y. L.; Xie, P. G.; Ma, J. G., Aberrant methylation of CDH13 is a potential biomarker for predicting the recurrence and progression of non muscle invasive bladder cancer. *Med. Sci. Monit* **2014,** *20*, 1572-1577.

338. Toyooka, K. O.; Toyooka, S.; Virmani, A. K.; Sathyanarayana, U. G.; Euhus, D. M.; Gilcrease, M.; Minna, J. D.; Gazdar, A. F., Loss of expression and aberrant methylation of the CDH13 (H-cadherin) gene in breast and lung carcinomas. *Cancer Res* **2001,** *61* (11), 4556-60.
339. Belinsky, S. A.; Palmisano, W. A.; Gilliland, F. D.; Crooks, L. A.; Divine, K. K.; Winters, S. A.; Grimes, M. J.; Harms, H. J.; Tellez, C. S.; Smith, T. M.; Moots, P. P.; Lechner, J. F.; Stidley, C. A.; Crowell, R. E., Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers. *Cancer Res* **2002,** *62* (8), 2370-2377.
340. Buckingham, L.; Penfield, F. L.; Kim, A.; Liptay, M.; Barger, C.; Basu, S.; Fidler, M.; Walters, K.; Bonomi, P.; Coon, J., PTEN, RASSF1 and DAPK site-specific hypermethylation and outcome in surgically treated stage I and II nonsmall cell lung cancer patients. *Int. J. Cancer* **2010,** *126* (7), 1630-1639.
341. Shenker, N. S.; Ueland, P. M.; Polidoro, S.; van Veldhoven, K.; Ricceri, F.; Brown, R.; Flanagan, J. M.; Vineis, P., DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology* **2013,** *24* (5), 712-716.
342. Philibert, R. A.; Beach, S. R.; Brody, G. H., Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics* **2012,** *7* (11), 1331-1338.
343. Tang, X.; Khuri, F. R.; Lee, J. J.; Kemp, B. L.; Liu, D.; Hong, W. K.; Mao, L., Hypermethylation of the death-associated protein (DAP) kinase promoter and aggressiveness in stage I non-small-cell lung cancer. *J Natl Cancer Inst* **2000,** *92* (18), 1511-6.
344. Chen, H.; Suzuki, M.; Nakamura, Y.; Ohira, M.; Ando, S.; Iida, T.; Nakajima, T.; Nakagawara, A.; Kimura, H., Aberrant methylation of RASGRF2 and RASSF1A in human non-small cell lung cancer. *Oncol Rep* **2006,** *15* (5), 1281-5.
345. Kaypee, S.; Sudarshan, D.; Shanmugam, M. K.; Mukherjee, D.; Sethi, G.; Kundu, T. K., Aberrant lysine acetylation in tumorigenesis: Implications in the development of therapeutics. *Pharmacol Ther* **2016,** *162*, 98-119.
346. Glozak, M. A.; Seto, E., Histone deacetylases and cancer. *Oncogene* **2007,** *26* (37), 5420-5432.
347. Aung, H. T.; Schroder, K.; Himes, S. R.; Brion, K.; van Zuylen, W.; Trieu, A.; Suzuki, H.; Hayashizaki, Y.; Hume, D. A.; Sweet, M. J.; Ravasi, T., LPS regulates proinflammatory gene expression in macrophages by altering histone deacetylase expression. *FASEB J* **2006,** *20* (9), 1315-27.
348. Licatalosi, D. D.; Darnell, R. B., RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet* **2010,** *11* (1), 75-87.
349. Lim, L. P.; Lau, N. C.; Garrett-Engele, P.; Grimson, A.; Schelter, J. M.; Castle, J.; Bartel, D. P.; Linsley, P. S.; Johnson, J. M., Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **2005,** *433* (7027), 769-773.

350. Gomes, M.; Teixeira, A. L.; Coelho, A.; Araujo, A.; Medeiros, R., The role of inflammation in lung cancer. *Adv Exp Med Biol* **2014,** *816*, 1-23.
351. Barkauskas, C. E.; Cronce, M. J.; Rackley, C. R.; Bowie, E. J.; Keene, D. R.; Stripp, B. R.; Randell, S. H.; Noble, P. W.; Hogan, B. L., Type 2 alveolar cells are stem cells in adult lung. *J Clin Invest* **2013,** *123* (7), 3025-36.
352. Prince, L. S.; Okoh, V. O.; Moninger, T. O.; Matalon, S., Lipopolysaccharide increases alveolar type II cell number in fetal mouse lungs through Toll-like receptor 4 and NF-kappaB. *Am J Physiol Lung Cell Mol Physiol* **2004,** *287* (5), L999-1006.
353. Song, C. X.; He, C., Potential functional roles of DNA demethylation intermediates. *Trends Biochem Sci* **2013,** *38* (10), 480-4.
354. John, G.; Kohse, K.; Orasche, J.; Reda, A.; Schnelle-Kreis, J.; Zimmermann, R.; Schmid, O.; Eickelberg, O.; Yildirim, A. O., The composition of cigarette smoke determines inflammatory cell recruitment to the lung in COPD mouse models. *Clin Sci (Lond)* **2014,** *126* (3), 207-21.
355. Rice, W. R.; Conkright, J. J.; Na, C. L.; Ikegami, M.; Shannon, J. M.; Weaver, T. E., Maintenance of the mouse type II cell phenotype in vitro. *Am J Physiol Lung Cell Mol Physiol* **2002,** *283* (2), L256-64.
356. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2007,** *2* (8), 1896-906.
357. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** *26* (12), 1367-1372.
358. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **2011,** *10* (4), 1794-1805.
359. Vazquez-Baeza, Y.; Pirrung, M.; Gonzalez, A.; Knight, R., EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* **2013,** *2* (1), 16.
360. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D. R.; Pimentel, H.; Salzberg, S. L.; Rinn, J. L.; Pachter, L., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **2012,** *7* (3), 562-78.
361. Mihelic, M.; Teuscher, C.; Turk, V.; Turk, D., Mouse stefins A1 and A2 (Stfa1 and Stfa2) differentiate between papain-like endo- and exopeptidases. *FEBS Lett* **2006,** *580* (17), 4195-9.
362. Djurec, M.; Grana, O.; Lee, A.; Troule, K.; Espinet, E.; Cabras, L.; Navas, C.; Blasco, M. T.; Martin-Diaz, L.; Burdiel, M.; Li, J.; Liu, Z.; Vallespinos, M.; Sanchez-Bueno, F.; Sprick, M. R.; Trumpp, A.; Sainz, B., Jr.; Al-Shahrour, F.; Rabadan, R.; Guerra, C.; Barbacid, M., Saa3 is a key mediator of the protumorigenic properties of cancer-associated fibroblasts in pancreatic tumors. *Proc Natl Acad Sci U S A* **2018,** *115* (6), E1147-E1156.

363. Loughner, C. L.; Bruford, E. A.; McAndrews, M. S.; Delp, E. E.; Swamynathan, S.; Swamynathan, S. K., Organization, evolution and functions of the human and mouse Ly6/uPAR family genes. *Hum Genomics* **2016,** *10*, 10.
364. Li, Z.; Su, D.; Ying, L.; Yu, G.; Mao, W., Study on expression of CDH4 in lung cancer. *World J Surg Oncol* **2017,** *15* (1), 26.
365. Du, C.; Huang, T.; Sun, D.; Mo, Y.; Feng, H.; Zhou, X.; Xiao, X.; Yu, N.; Hou, B.; Huang, G.; Ernberg, I.; Zhang, Z., CDH4 as a novel putative tumor suppressor gene epigenetically silenced by promoter hypermethylation in nasopharyngeal carcinoma. *Cancer Lett* **2011,** *309* (1), 54-61.
366. Lee, C. G.; Da Silva, C. A.; Dela Cruz, C. S.; Ahangari, F.; Ma, B.; Kang, M. J.; He, C. H.; Takyar, S.; Elias, J. A., Role of chitin and chitinase/chitinase-like proteins in inflammation, tissue remodeling, and injury. *Annu Rev Physiol* **2011,** *73*, 479-501.
367. Tan, W.; Zolotukhin, A. S.; Tretyakova, I.; Bear, J.; Lindtner, S.; Smulevitch, S. V.; Felber, B. K., Identification and characterization of the mouse nuclear export factor (Nxf) family members. *Nucleic Acids Res* **2005,** *33* (12), 3855-65.
368. Zhou, X.; D'Agostino, J.; Xie, F.; Ding, X., Role of CYP2A5 in the bioactivation of the lung carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone in mice. *J Pharmacol Exp Ther* **2012,** *341* (1), 233-41.
369. Zergoun, A. A.; Zebboudj, A.; Sellam, S. L.; Kariche, N.; Djennaoui, D.; Ouraghi, S.; Kerboua, E.; Amir-Tidadini, Z. C.; Chilla, D.; Asselah, F.; Touil-Boukoffa, C.; Merghoub, T.; Bourouba, M., IL-6/NOS2 inflammatory signals regulate MMP-9 and MMP-2 activity and disease outcome in nasopharyngeal carcinoma patients. *Tumour Biol* **2016,** *37* (3), 3505-14.
370. Mehner, C.; Hockla, A.; Miller, E.; Ran, S.; Radisky, D. C.; Radisky, E. S., Tumor cell-produced matrix metalloproteinase 9 (MMP-9) drives malignant progression and metastasis of basal-like triple negative breast cancer. *Oncotarget* **2014,** *5* (9), 2736-49.
371. Shankaran, V.; Ikeda, H.; Bruce, A. T.; White, J. M.; Swanson, P. E.; Old, L. J.; Schreiber, R. D., IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **2001,** *410* (6832), 1107-11.
372. Lu, J.; McCarter, M.; Lian, G.; Esposito, G.; Capoccia, E.; Delli-Bovi, L. C.; Hecht, J.; Sheen, V., Global hypermethylation in fetal cortex of Down syndrome due to DNMT3L overexpression. *Hum Mol Genet* **2016,** *25* (9), 1714-27.
373. Drilon, A.; Sugita, H.; Sima, C. S.; Zauderer, M.; Rudin, C. M.; Kris, M. G.; Rusch, V. W.; Azzoli, C. G., A prospective study of tumor suppressor gene methylation as a prognostic biomarker in surgically resected stage I to IIIA non-small-cell lung cancers. *J Thorac Oncol* **2014,** *9* (9), 1272-7.
374. Davis, M. R.; Zhu, Z.; Hansen, D. M.; Bai, Q.; Fang, Y., The role of IL-21 in immunity and cancer. *Cancer Lett* **2015,** *358* (2), 107-114.
375. Brody, J. S.; Spira, A., State of the art. Chronic obstructive pulmonary disease, inflammation, and lung cancer. *Proc Am Thorac Soc* **2006,** *3* (6), 535-7.
376. Burn, J.; Gerdes, A. M.; Macrae, F.; Mecklin, J. P.; Moeslein, G.; Olschwang, S.; Eccles, D.; Evans, D. G.; Maher, E. R.; Bertario, L.; Bisgaard, M. L.; Dunlop, M. G.; Ho,

J. W.; Hodgson, S. V.; Lindblom, A.; Lubinski, J.; Morrison, P. J.; Murday, V.; Ramesar, R.; Side, L.; Scott, R. J.; Thomas, H. J.; Vasen, H. F.; Barker, G.; Crawford, G.; Elliott, F.; Movahedi, M.; Pylvanainen, K.; Wijnen, J. T.; Fodde, R.; Lynch, H. T.; Mathers, J. C.; Bishop, D. T.; Investigators, C., Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *Lancet* **2011,** *378* (9809), 2081-7.
377. Talikka, M.; Sierro, N.; Ivanov, N. V.; Chaudhary, N.; Peck, M. J.; Hoeng, J.; Coggins, C. R.; Peitsch, M. C., Genomic impact of cigarette smoke, with application to three smoking-related diseases. *Crit Rev. Toxicol* **2012,** *42* (10), 877-889.
378. Ehrlich, M.; Ehrlich, K. C., DNA cytosine methylation and hydroxymethylation at the borders. *Epigenomics* **2014,** *6* (6), 563-6.
379. McAlister, G. C.; Nusinow, D. P.; Jedrychowski, M. P.; Wuhr, M.; Huttlin, E. L.; Erickson, B. K.; Rad, R.; Haas, W.; Gygi, S. P., MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem* **2014,** *86* (14), 7150-7158.
380. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D., The human genome browser at UCSC. *Genome Res* **2002,** *12* (6), 996-1006.
381. UniProt Consortium, T., UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2018,** *46* (5), 2699.
382. Huang da, W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4* (1), 44-57.
383. Huang da, W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009,** *37* (1), 1-13.
384. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nat Methods* **2009,** *6* (5), 359-62.
385. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **1996,** *68* (5), 850-8.
386. Bird, A., DNA methylation patterns and epigenetic memory. *Genes Dev* **2002,** *16* (1), 6-21.
387. Baeza-Loya, S.; Viswanath, H.; Carter, A.; Molfese, D. L.; Velasquez, K. M.; Baldwin, P. R.; Thompson-Lake, D. G.; Sharp, C.; Fowler, J. C.; De La Garza, R., 2nd; Salas, R., Perceptions about e-cigarette safety may lead to e-smoking during pregnancy. *Bull Menninger Clin* **2014,** *78* (3), 243-52.
388. Lee, H. W.; Park, S. H.; Weng, M. W.; Wang, H. T.; Huang, W. C.; Lepor, H.; Wu, X. R.; Chen, L. C.; Tang, M. S., E-cigarette smoke damages DNA and reduces repair activity in mouse lung, heart, and bladder as well as in human lung and bladder cells. *Proc Natl Acad Sci U S A* **2018,** *115* (7), E1560-E1569.
389. Martin, E. M.; Clapp, P. W.; Rebuli, M. E.; Pawlak, E. A.; Glista-Baker, E.; Benowitz, N. L.; Fry, R. C.; Jaspers, I., E-cigarette use results in suppression of immune

and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke. *Am J Physiol Lung Cell Mol Physiol* **2016,** *311* (1), L135-44.
390. Allen, J. G.; Flanigan, S. S.; LeBlanc, M.; Vallarino, J.; MacNaughton, P.; Stewart, J. H.; Christiani, D. C., Flavoring Chemicals in E-Cigarettes: Diacetyl, 2,3-Pentanedione, and Acetoin in a Sample of 51 Products, Including Fruit-, Candy-, and Cocktail-Flavored E-Cigarettes. *Environ Health Perspect* **2016,** *124* (6), 733-9.
391. Geiss, O.; Bianchi, I.; Barrero-Moreno, J., Correlation of volatile carbonyl yields emitted by e-cigarettes with the temperature of the heating coil and the perceived sensorial quality of the generated vapours. *Int J Hyg Environ Health* **2016,** *219* (3), 268-77.
392. Gillman, I. G.; Kistler, K. A.; Stewart, E. W.; Paolantonio, A. R., Effect of variable power levels on the yield of total aerosol mass and formation of aldehydes in e-cigarette aerosols. *Regul Toxicol Pharmacol* **2016,** *75*, 58-65.
393. Holden, V. K.; Hines, S. E., Update on flavoring-induced lung disease. *Curr Opin Pulm Med* **2016,** *22* (2), 158-64.
394. Chen, C.; Huang, H.; Wu, C. H., Protein Bioinformatics Databases and Resources. *Methods Mol Biol* **2017,** *1558*, 3-39.
395. He, Y.; Smith, R., Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. *Cell Mol Life Sci* **2009,** *66* (7), 1239-56.
396. Munro, T. P.; Magee, R. J.; Kidd, G. J.; Carson, J. H.; Barbarese, E.; Smith, L. M.; Smith, R., Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking. *J Biol Chem* **1999,** *274* (48), 34389-95.
397. Alarcon, C. R.; Goodarzi, H.; Lee, H.; Liu, X.; Tavazoie, S.; Tavazoie, S. F., HNRNPA2B1 Is a Mediator of m(6)A-Dependent Nuclear RNA Processing Events. *Cell* **2015,** *162* (6), 1299-308.
398. Lin, Y. L.; Shivji, M. K.; Chen, C.; Kolodner, R.; Wood, R. D.; Dutta, A., The evolutionarily conserved zinc finger motif in the largest subunit of human replication protein A is required for DNA replication and mismatch repair but not for nucleotide excision repair. *J Biol Chem* **1998,** *273* (3), 1453-61.
399. Shykind, B. M.; Kim, J.; Sharp, P. A., Activation of the TFIID-TFIIA complex with HMG-2. *Genes Dev* **1995,** *9* (11), 1354-65.
400. Fan, Z.; Beresford, P. J.; Zhang, D.; Lieberman, J., HMG2 interacts with the nucleosome assembly protein SET and is a target of the cytotoxic T-lymphocyte protease granzyme A. *Mol Cell Biol* **2002,** *22* (8), 2810-20.
401. Ugrinova, I.; Pashev, I. G.; Pasheva, E. A., Nucleosome binding properties and Co-remodeling activities of native and in vivo acetylated HMGB-1 and HMGB-2 proteins. *Biochemistry* **2009,** *48* (27), 6502-7.

APPENDICES

A1. Chapter II: NMR Spectra

Figure A1.1. ^{1}H-NMR of 5-iodo-2′-deoxycytidine (**1**).

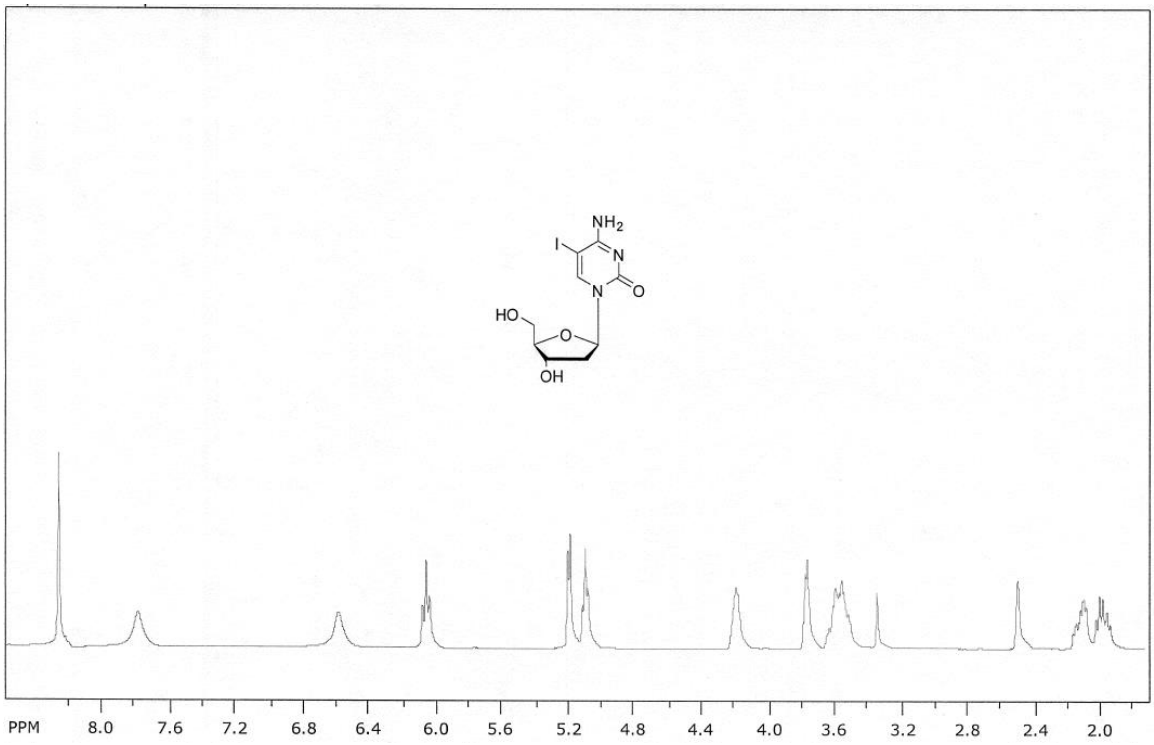


Figure A1.2. ^{1}H-NMR of 5-ethyl-3′,5′-O-t-butyldimethylsilyl-2′-deoxycytidine (**3**).

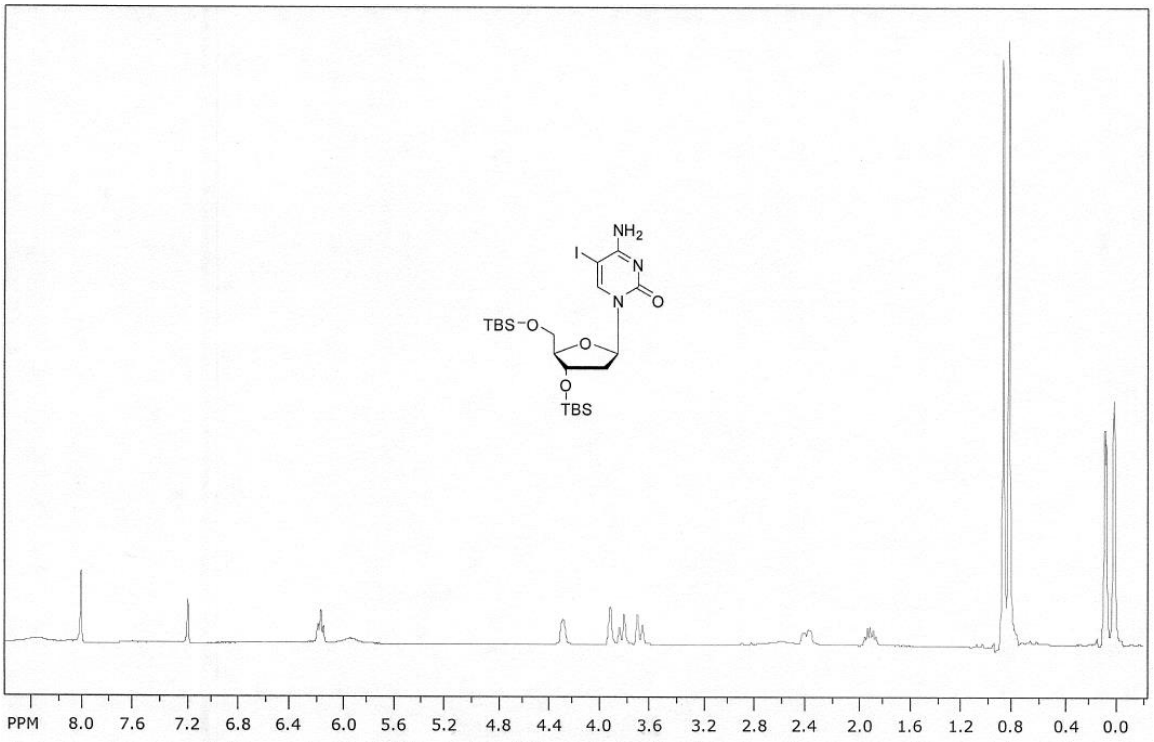


Figure A1.3. ^{1}H-NMR of 5-(Trimethylsilyl)-ethynyl-3′,5′-O-t-butyldimethylsilyl-2′-deoxycytidine (**3**).

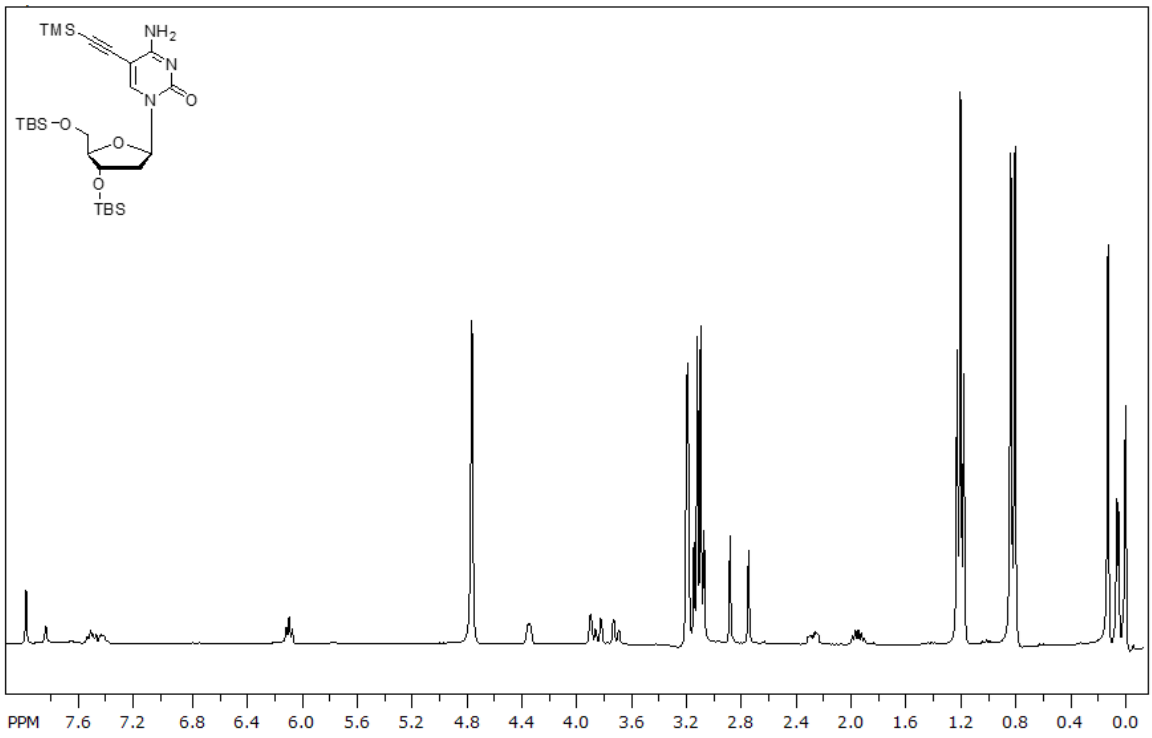


Figure A1.4. ^{1}H-NMR of 5-Ethynyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**4**).

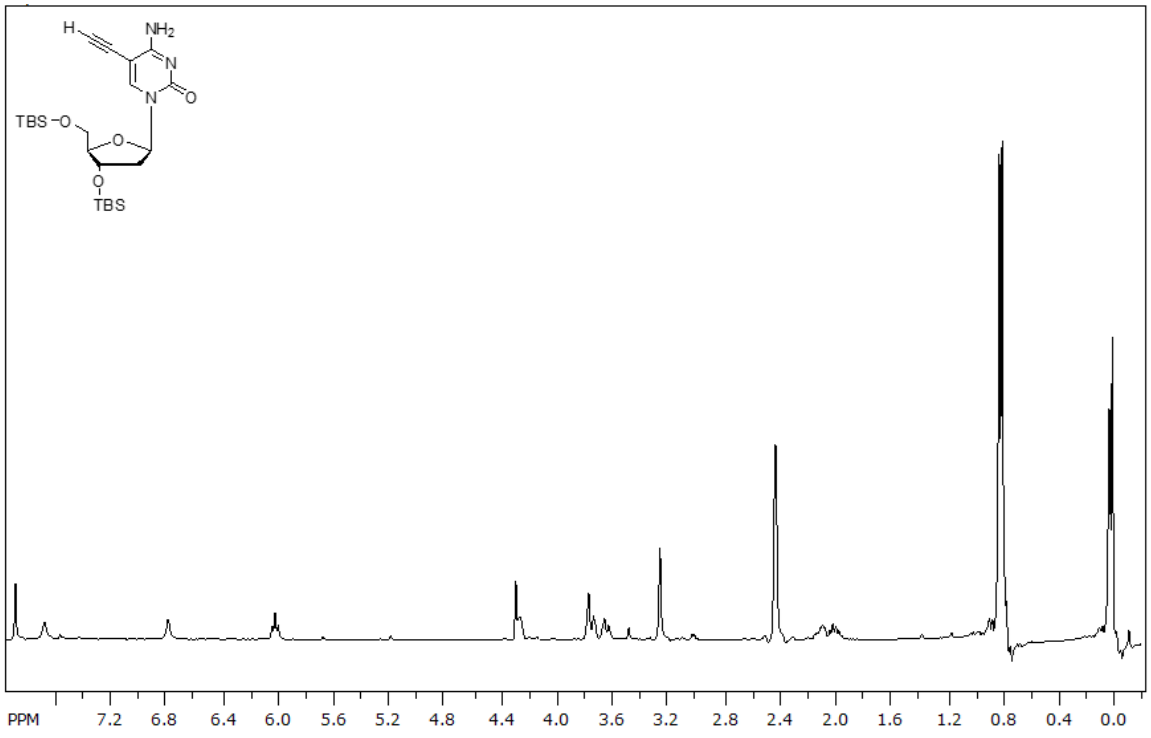


Figure A1.5. ^{1}H-NMR of 5-Ethyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**5**).

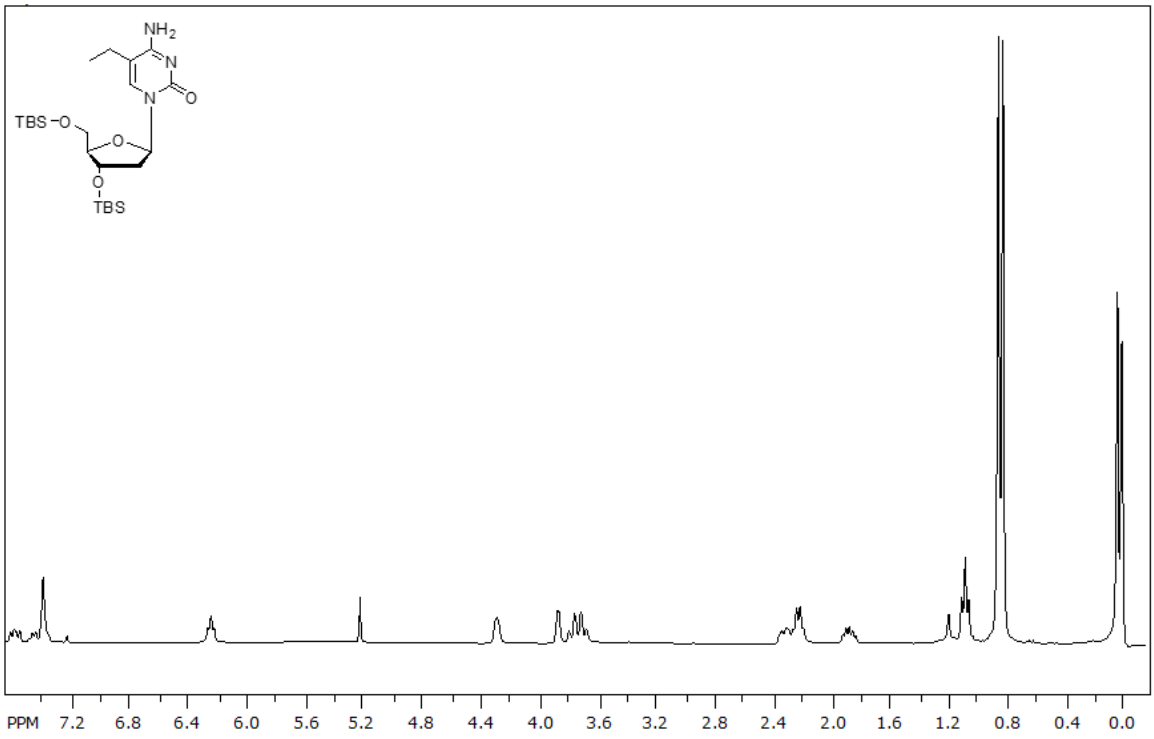


Figure A1.6. ^{1}H-NMR of 4-*N*-Benzoyl-5-ethyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**6**).

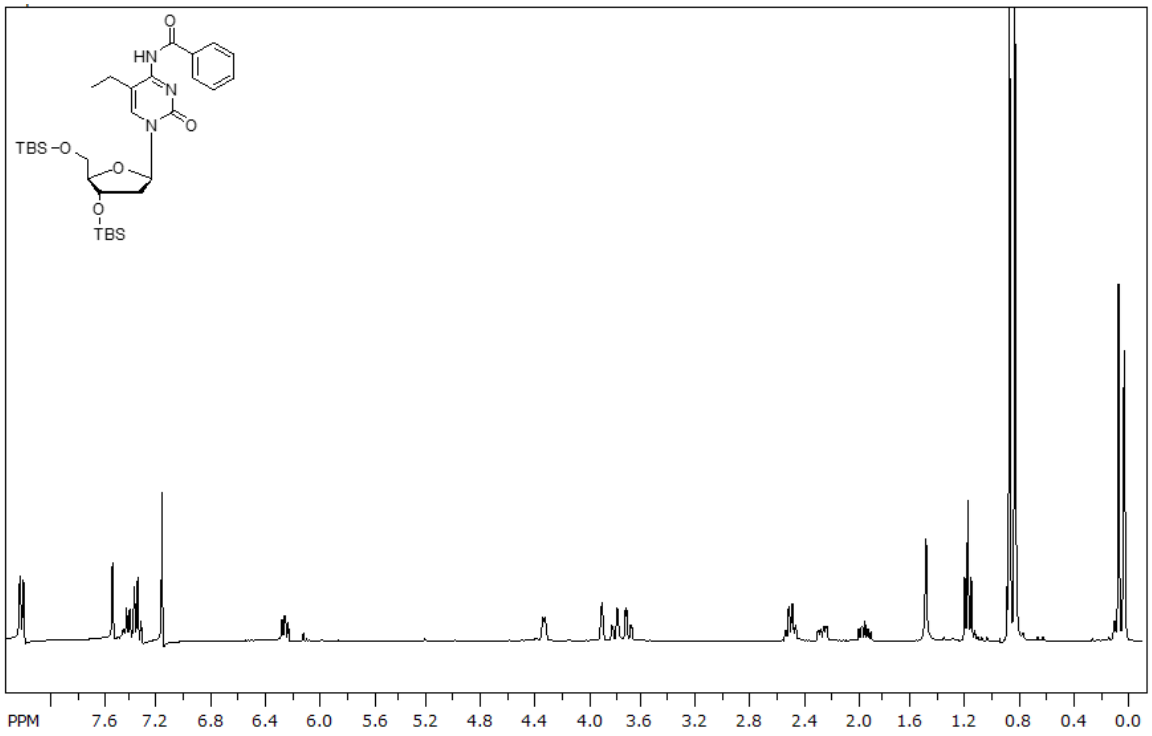


Figure A1.7. ^{1}H-NMR of 4-*N*-Benzoyl-5-ethyl-2′-deoxycytidine (**7**).

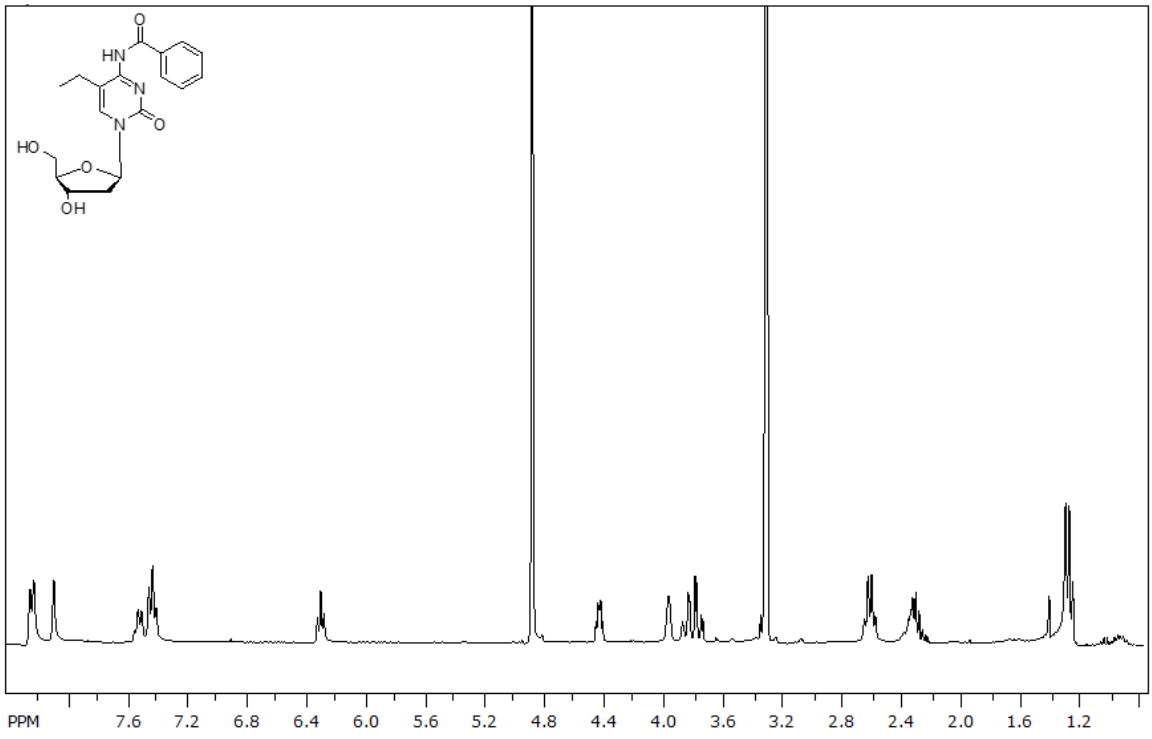


Figure A1.8. ^{1}H-NMR of 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-ethyl-2′-deoxycytidine (**8**).

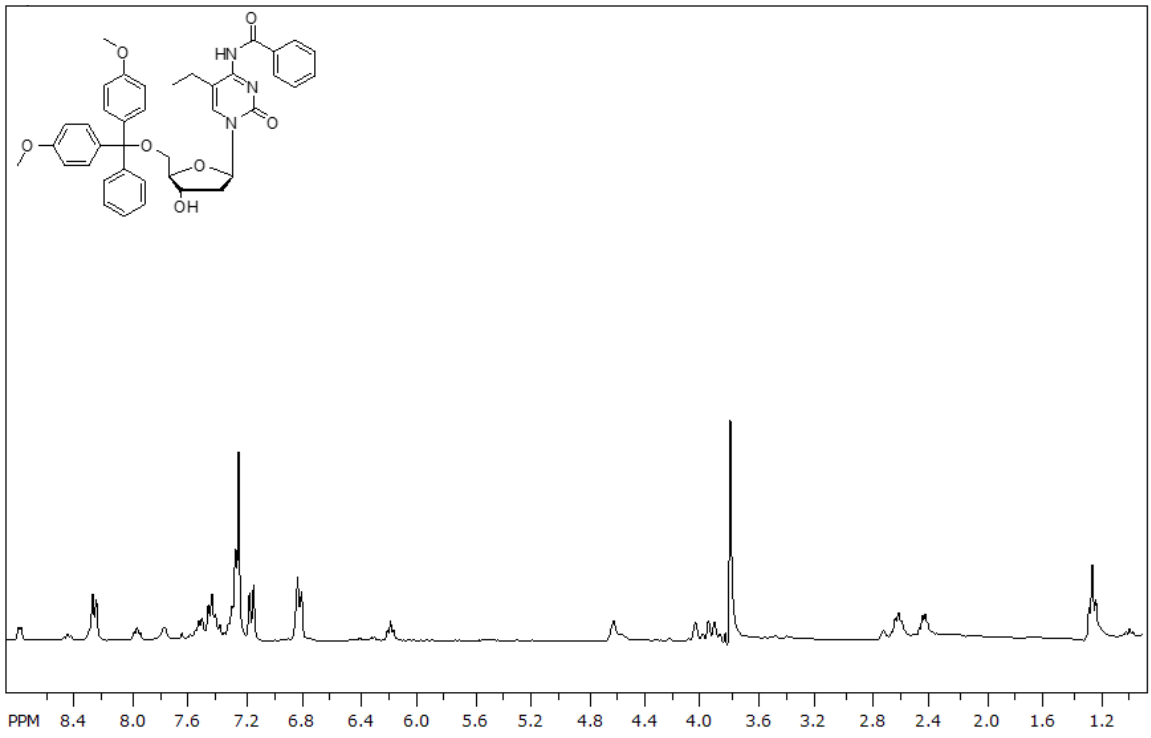


Figure A1.9. ^{31}P-NMR of 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-ethyl-2-deoxycytidine-3′-[(2-cyanoethyl)-(*N,N*-diisopropyl)]-phosphoramidite (**9**).

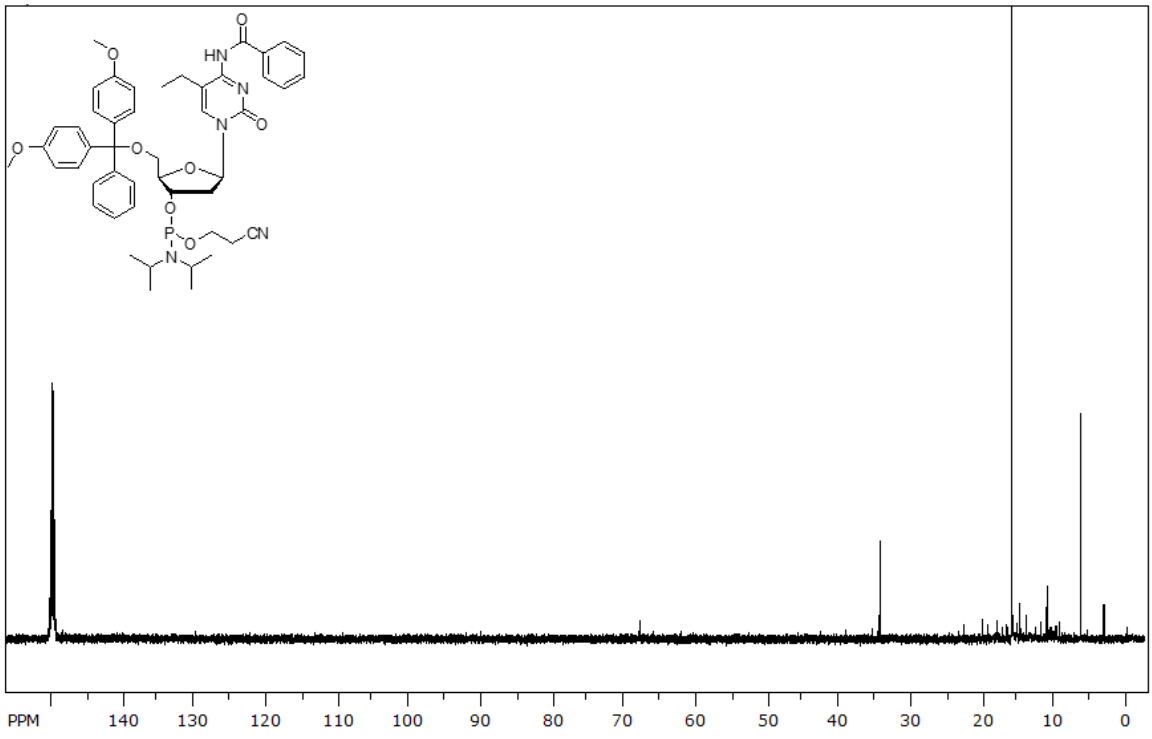


Figure A1.10. ^{1}H-NMR of 5-Propyn-1-yl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**10**).

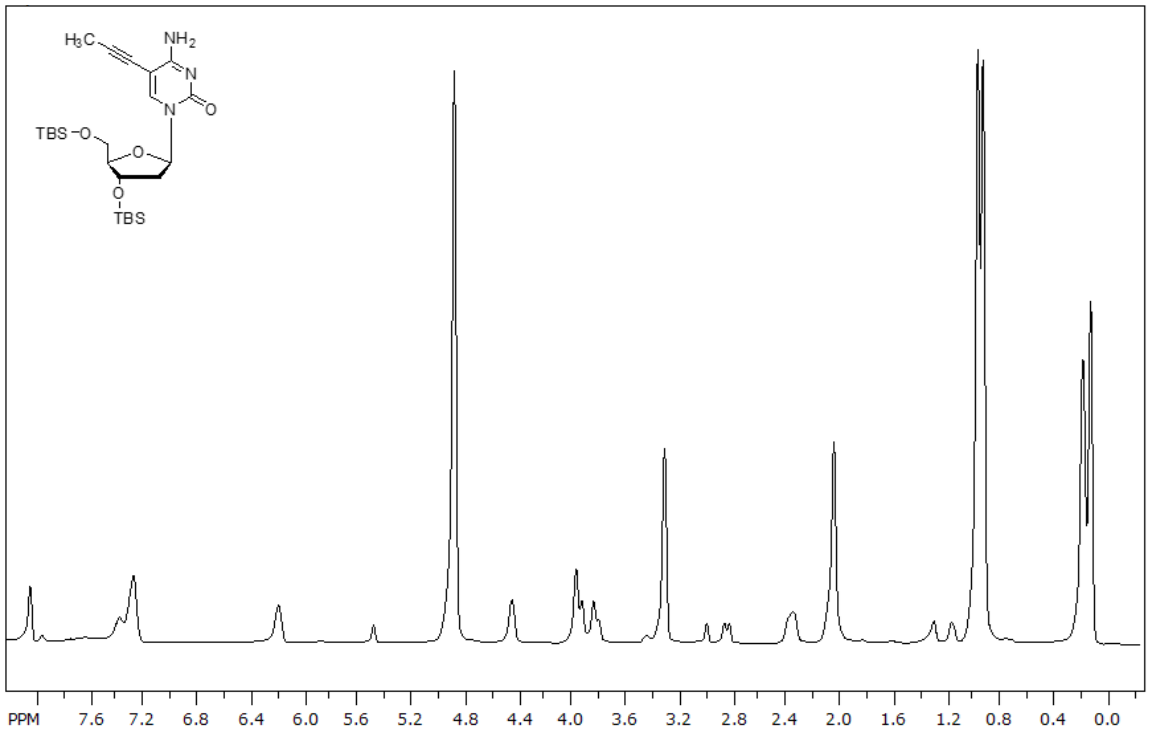


Figure A1.11. ^{1}H-NMR and COSY of 5-Propyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**11**).

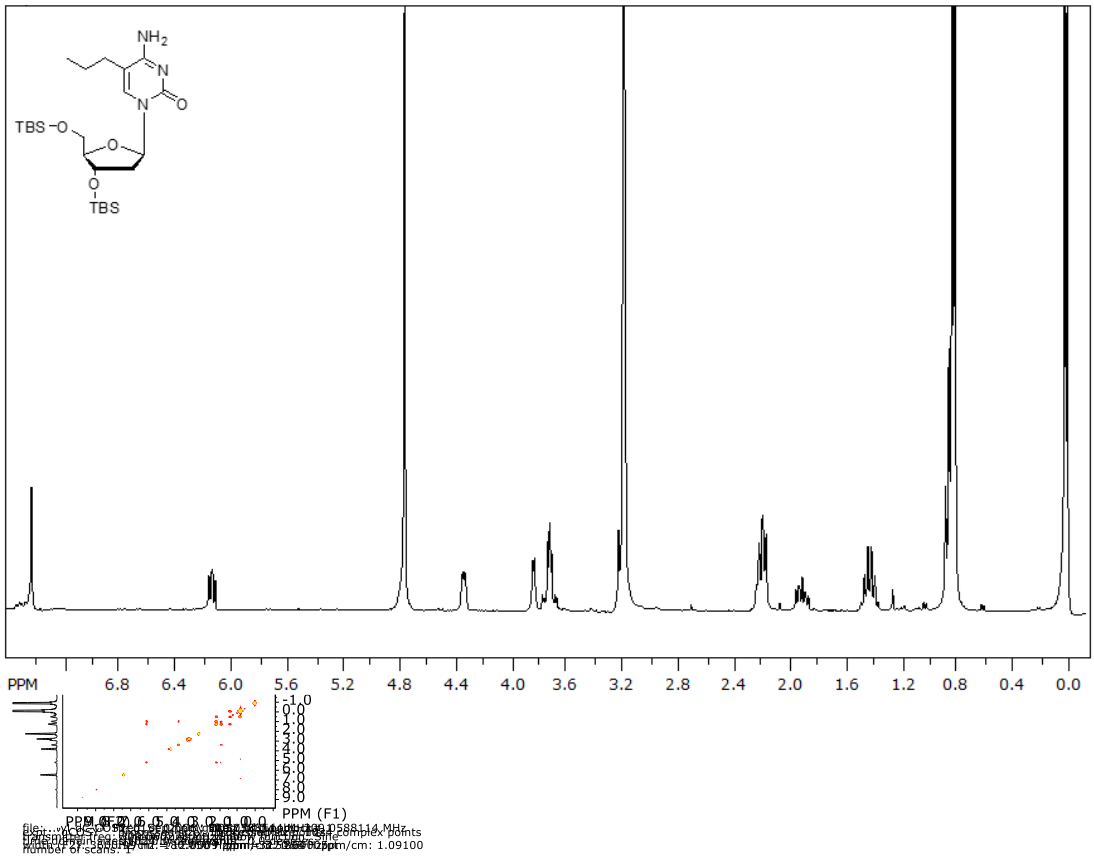


Figure A1.12. ^{1}H-NMR of 4-*N*-Benzoyl-5-propyl-3′,5′-*O*-*t*-butyldimethylsilyl-2′-deoxycytidine (**12**).

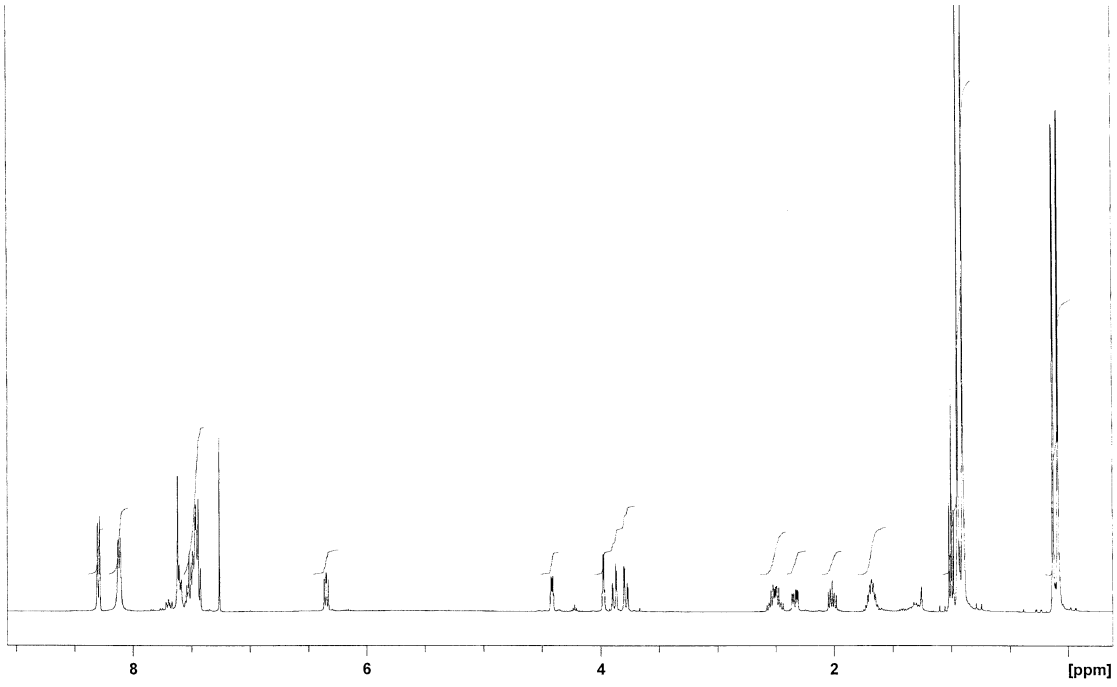


Figure A1.13. ^{1}H-NMR of 4-*N*-Benzoyl-5-propyl-2′-deoxycytidine (**13**).

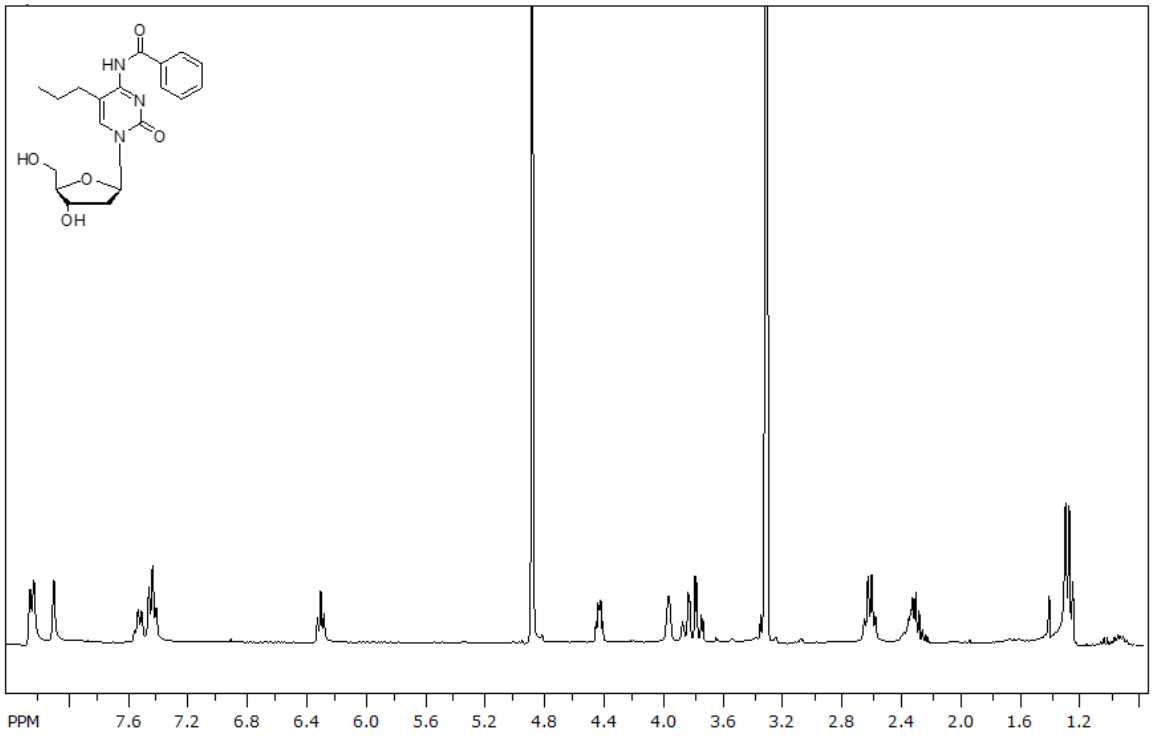


Figure A1.14. ^{1}H-NMR of 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-propyl-2′-deoxycytidine (**14**).

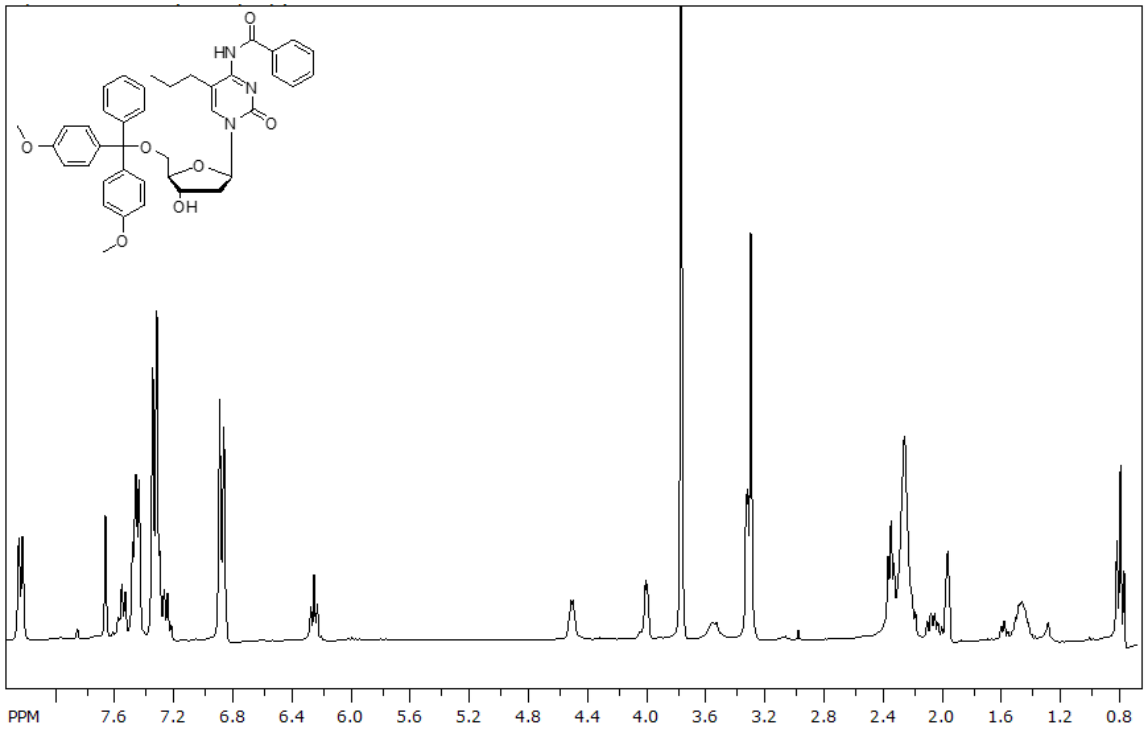


Figure A1.15. ^{1}H-NMR of 3′,5′-*O*-*t*-butyldimethylsilyl-5-vinyl-2′-deoxycytidine (**16**).

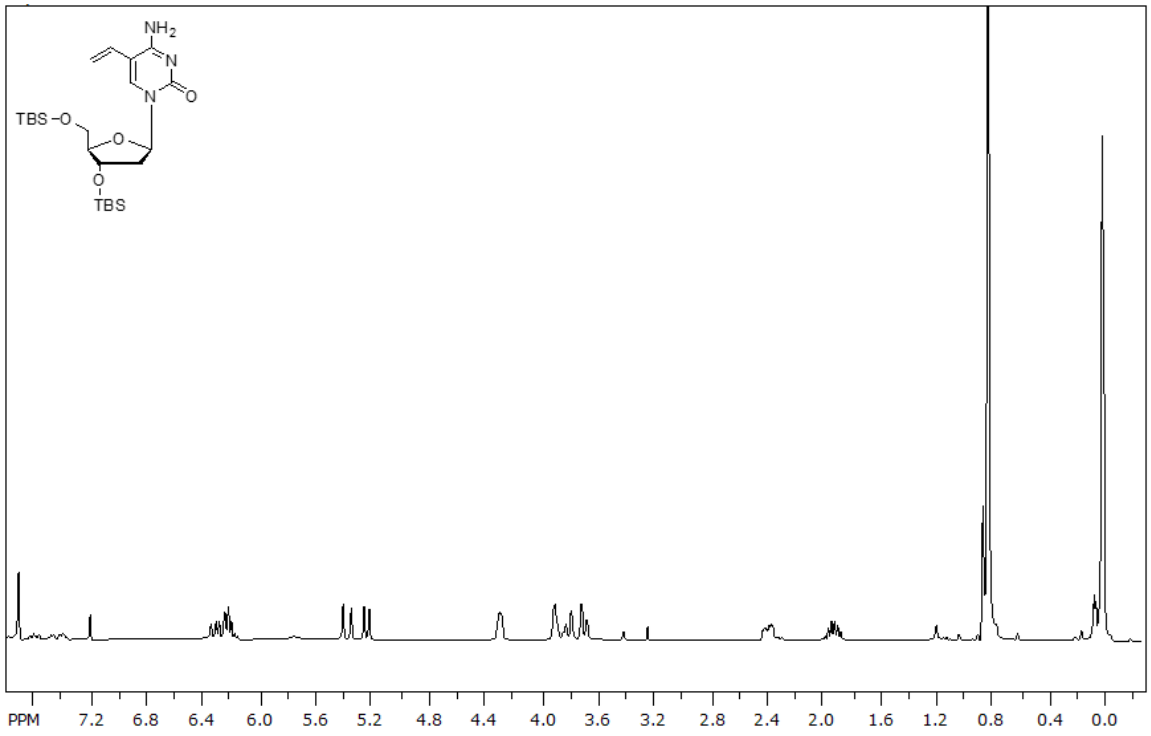


Figure A1.16. ^{1}H-NMR of 4-*N*-Benzoyl-3′,5′-*O*-(*t*-butyldimethylsilyl)- 5-vinyl-2′-deoxycytidine (**17**).

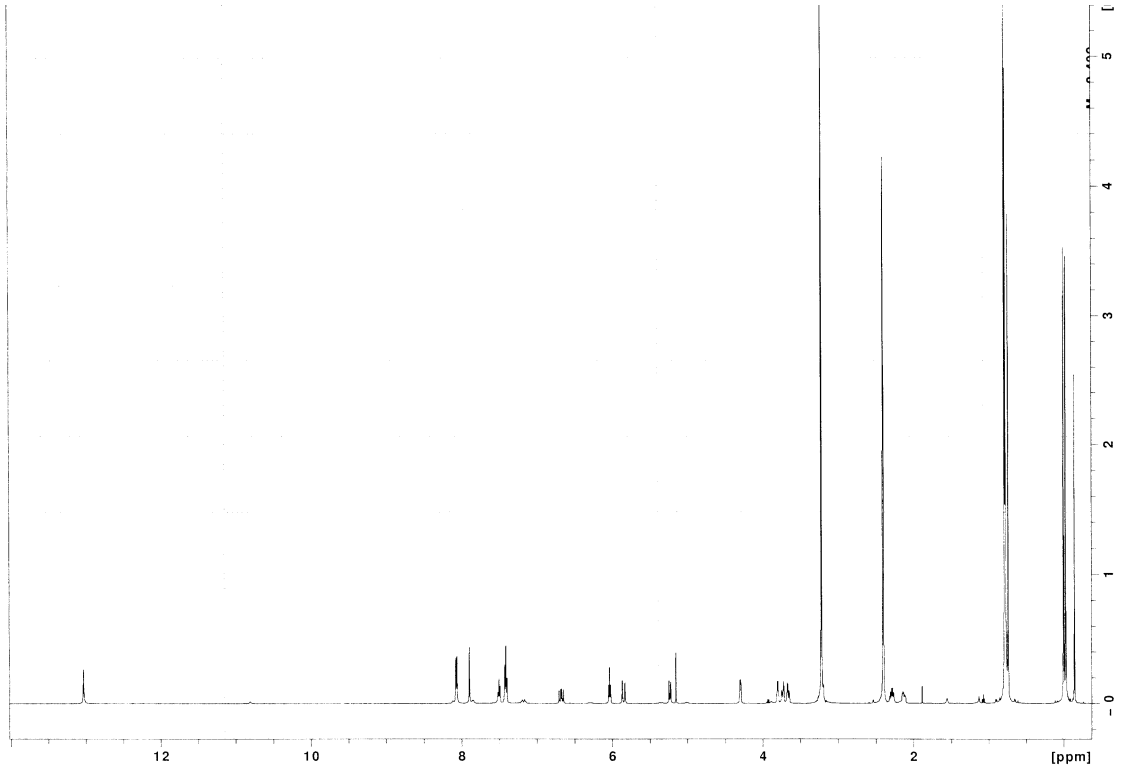


Figure A1.17. [1]H-NMR and COSY of 4-*N*-Benzoyl-5-vinyl-2′-deoxycytidine (**18**).

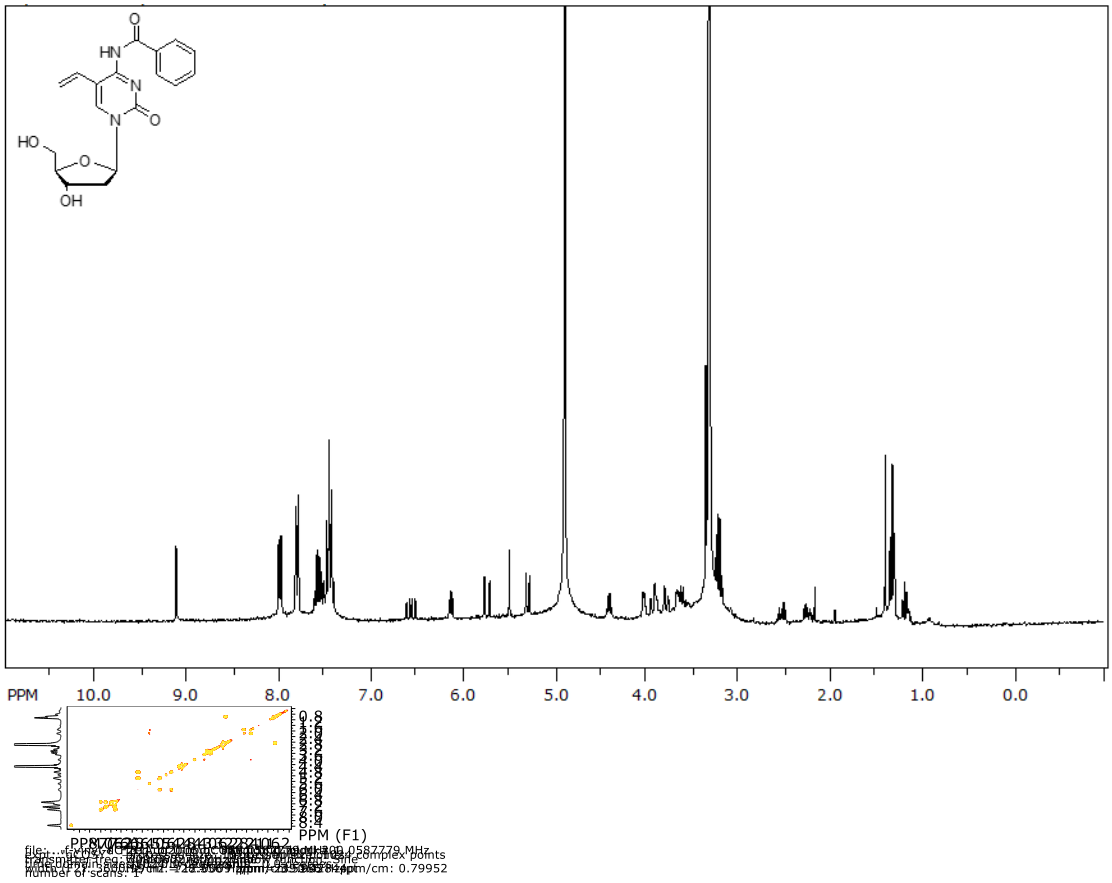


Figure A1.18. ^{1}H-NMR of 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-vinyl- 2′deoxycytidine (**19**).

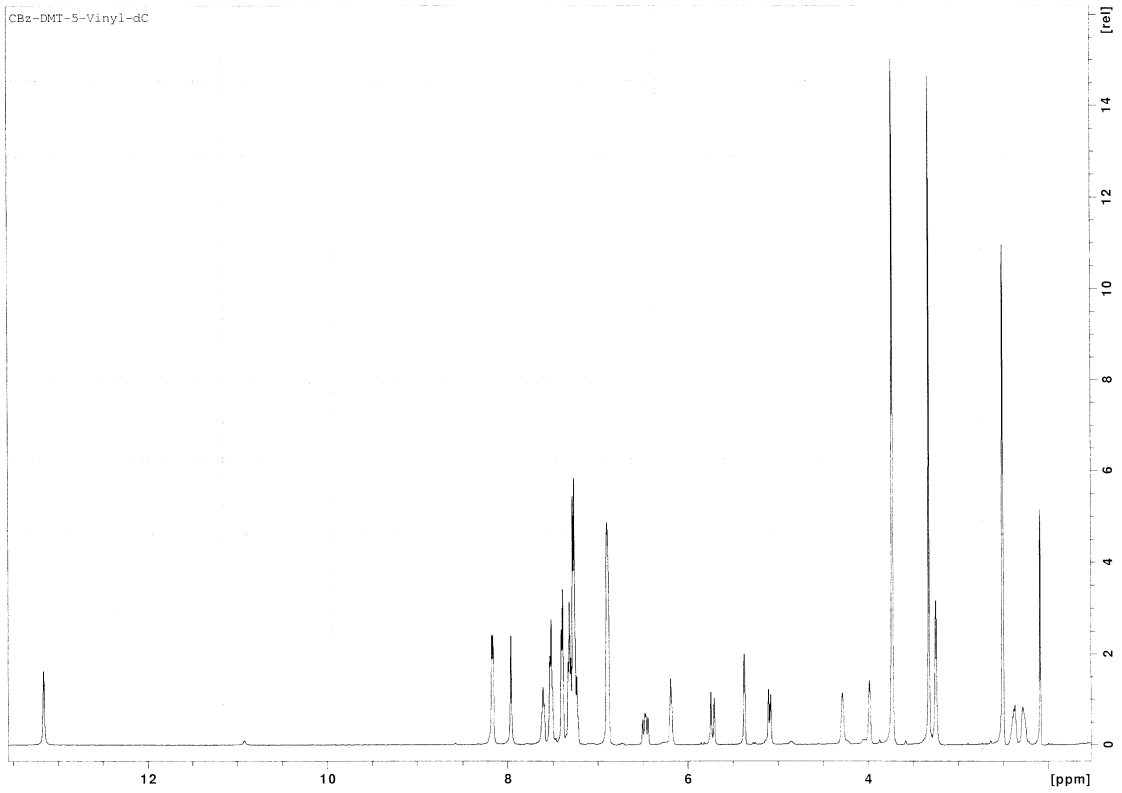
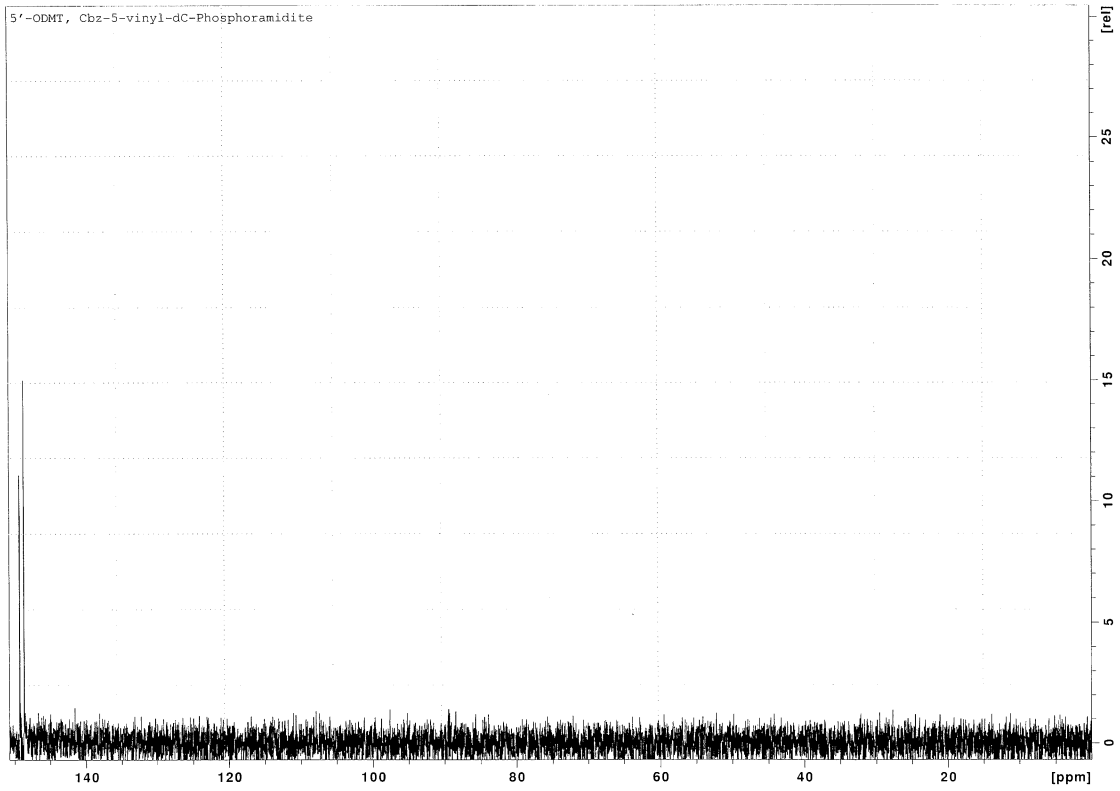
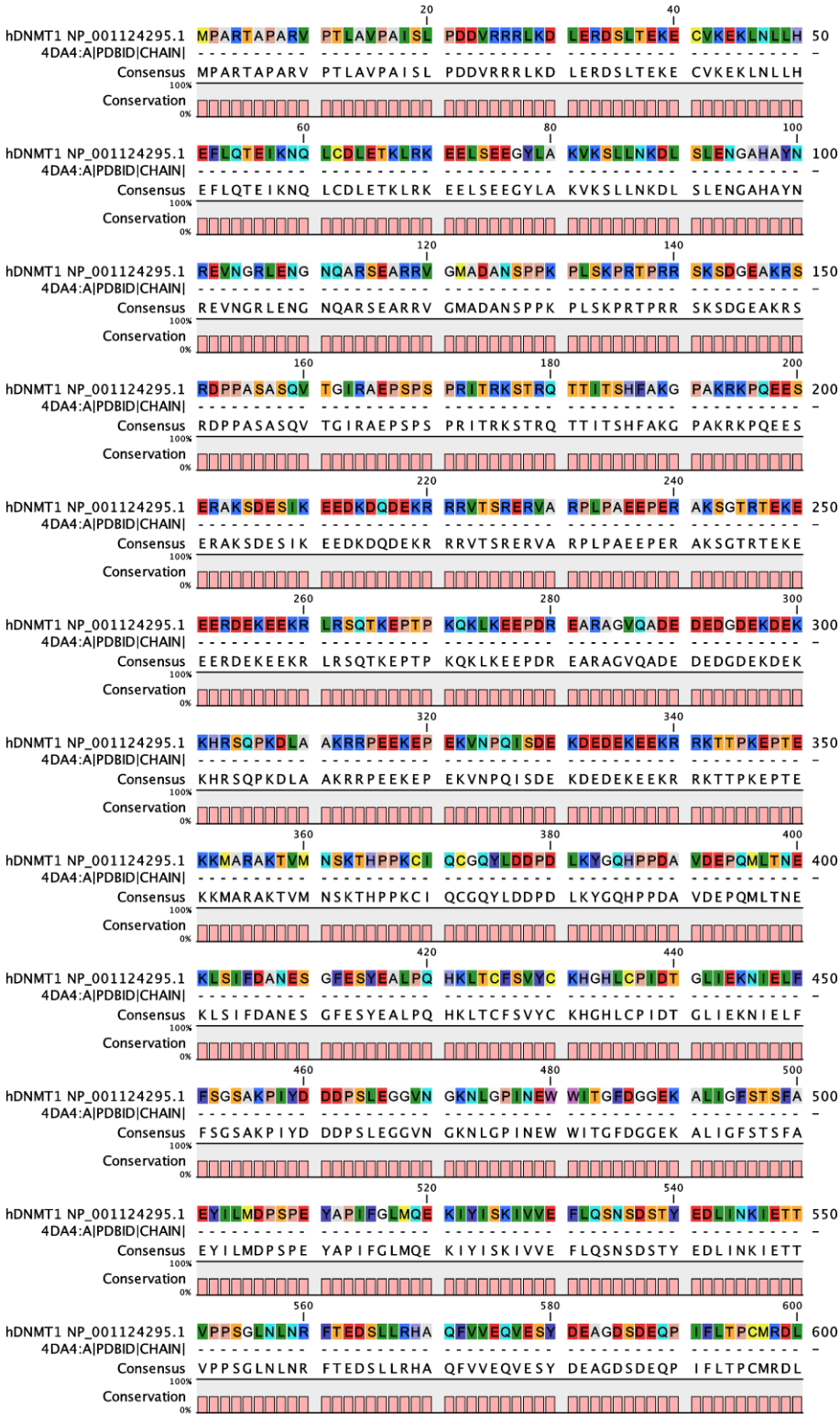


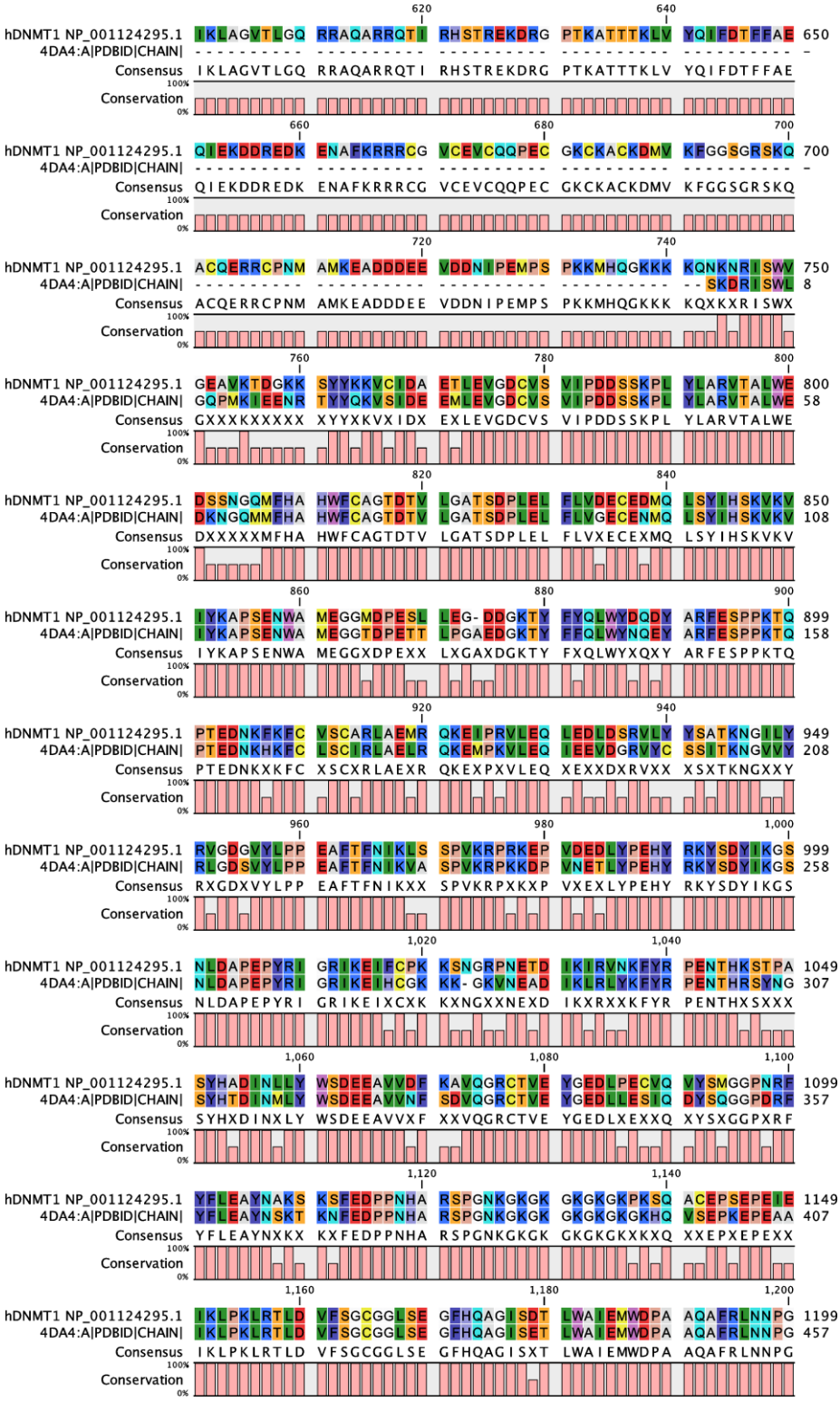
Figure A1.19. ^{31}P-NMR of 4-*N*-Benzoyl-5′-*O*-(dimethoxytrityl)-5-vinyl-2′deoxycytidine 3′-(2-cyanoethyl)-*N*, *N*′-diisopropyl phosphoramidite (**20**).

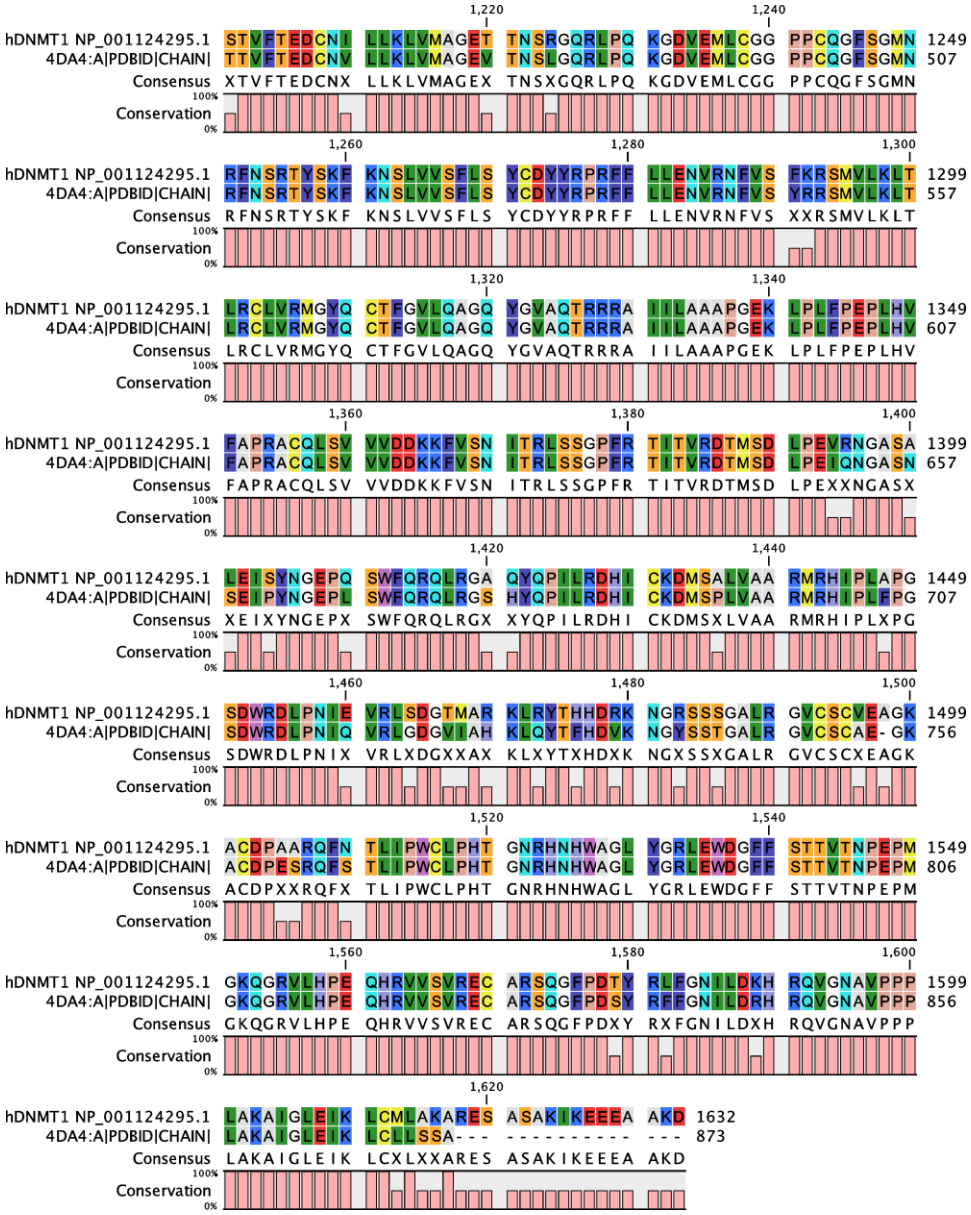


A2. Chapter III: Homology Model

Figure A2.1: Sequence alignment for hDNMT1 homology model. Modeling of hDNMT1 was carried out using the published crystal structure of mDNMT1 (PDB: 4DA4) in complex with hemi-methylated DNA[10] with the hDNMT1 reference sequence (NP_001124295.1). Alignment of the hDNMT1 reference sequence with the sequence of the mDNMT1 crystal structure reveals an 85% identity.


hDNMT1 NP_001124295.1
4DA4:A|PDBID|CHAIN|
Consensus
Conservation
100%
0%
MPARTAPARV PTLAVPAISL PDDVRRRLKD LERDSLTEKE CVKEKLNLLH 50
EFLQTEIKNQ LCDLETKLRK EELSEEGYLA KVKSLLNKDL SLENGAHAYN 100
REVNGRLENG NQARSEARRV GMADANSPPK PLSKPRTPRR SKSDGEAKRS 150
RDPPASASQV TGIRAEPSPS PRITRKSTRQ TTITSHFAKG PAKRKPQEES 200
ERAKSDESIK EEDKDQDEKR RRVTSRERVA RPLPAEEPER AKSGTRTEKE 250
EERDEKEEKR LRSQTKEPTP KQKLKEEPDR EARAGVQADE DEDGDEKDEK 300
KHRSQPKDLA AKRRPEEKEP EKVNPQISDE KDEDEKEEKR RKTTPKEPTE 350
KKMARAKTVM NSKTHPPKCI QCGQYLDDPD LKYGQHPPDA VDEPQMLTNE 400
KLSIFDANES GFESYEALPQ HKLTCFSVYC KHGHLCPIDT GLIEKNIELF 450
FSGSAKPIYD DDPSLEGGVN GKNLGPINEW WITGFDGGEK ALIGFSTSFA 500
EYILMDPSPE YAPIFGLMQE KIYISKIVVE FLQSNSDSTY EDLINKIETT 550
VPPSGLNLNR FTEDSLLRHA QFVVEQVESY DEAGDSDEQP IFLTPCMRDL 600


hDNMT1 NP_001124295.1 IKLAGVTLGQ RRAQARRQTI RHSTREKDRG PTKATTTKLV YQIFDTFFAE 650
4DA4:A|PDBID|CHAIN| ---------- ---------- ---------- ---------- ---------- -
Consensus IKLAGVTLGQ RRAQARRQTI RHSTREKDRG PTKATTTKLV YQIFDTFFAE
Conservation
hDNMT1 NP_001124295.1 QIEKDDREDK ENAFKRRRCG VCEVCQQPEC GKCKACKDMV KFGGSGRSKQ 700
4DA4:A|PDBID|CHAIN| ---------- ---------- ---------- ---------- ---------- -
Consensus QIEKDDREDK ENAFKRRRCG VCEVCQQPEC GKCKACKDMV KFGGSGRSKQ
Conservation
hDNMT1 NP_001124295.1 ACQERRCPNM AMKEADDDEE VDDNIPEMPS PKKMHQGKKK KQNKNRISWV 750
4DA4:A|PDBID|CHAIN| ---------- ---------- ---------- ---------- --SKDRISWL 8
Consensus ACQERRCPNM AMKEADDDEE VDDNIPEMPS PKKMHQGKKK KQXKXRISWX
Conservation
hDNMT1 NP_001124295.1 GEAVKTDGKK SYYKKVCIDA ETLEVGDCVS VIPDDSSKPL YLARVTALWE 800
4DA4:A|PDBID|CHAIN| GQPMKIEENR TYYQKVSIDE EMLEVGDCVS VIPDDSSKPL YLARVTALWE 58
Consensus GXXXKXXXXX XYYXKVXIDX EXLEVGDCVS VIPDDSSKPL YLARVTALWE
Conservation
hDNMT1 NP_001124295.1 DSSNGQMFHA HWFCAGTDTV LGATSDPLEL FLVDECEDMQ LSYIHSKVKV 850
4DA4:A|PDBID|CHAIN| DKNGQMMFHA HWFCAGTDTV LGATSDPLEL FLVGECENMQ LSYIHSKVKV 108
Consensus DXXXXXMFHA HWFCAGTDTV LGATSDPLEL FLVXECEXMQ LSYIHSKVKV
Conservation
hDNMT1 NP_001124295.1 IYKAPSENWA MEGGMDPESL LEG-DDGKTY FYQLWYDQDY ARFESPPKTQ 899
4DA4:A|PDBID|CHAIN| IYKAPSENWA MEGGTDPETT LPGAEDGKTY FEQLWYNQEY ARFESPPKTQ 158
Consensus IYKAPSENWA MEGGXDPEXX LXGAXDGKTY FXQLWYXQXY ARFESPPKTQ
Conservation
hDNMT1 NP_001124295.1 PTEDNKFKFC VSCARLAEMR QKEIPRVLEQ LEDLDSRVLY YSATKNGILY 949
4DA4:A|PDBID|CHAIN| PTEDNKHKFC LSCIRLAELR QKEMPKVLEQ IEEVDGRVYC SSITKNGVVY 208
Consensus PTEDNKXKFC XSCXRLAEXR QKEXPXVLEQ XEXXDXRVXX XSXTKNGXXY
Conservation
hDNMT1 NP_001124295.1 RVGDGVYLPP EAFTFNIKLS SPVKRPRKEP VDEDLYPEHY RKYSDYIKGS 999
4DA4:A|PDBID|CHAIN| RLGDSVYLPP EAFTFNIKVA SPVKRPKKDP VNETLYPEHY RKYSDYIKGS 258
Consensus RXGDXVYLPP EAFTFNIKXX SPVKRPXKXP VXEXLYPEHY RKYSDYIKGS
Conservation
hDNMT1 NP_001124295.1 NLDAPEPYRI GRIKEIFCPK KSNGRPNETD IKIRVNKFYR PENTHKSTPA 1049
4DA4:A|PDBID|CHAIN| NLDAPEPYRI GRIKEIHCGK KK-GKVNEAD IKLRLYKFYR PENTHRSYNG 307
Consensus NLDAPEPYRI GRIKEIXCXK KXNGXXNEXD IKXRXXKFYR PENTHXSXXX
Conservation
hDNMT1 NP_001124295.1 SYHADINLLY WSDEEAVVDF KAVQGRCTVE YGEDLPECVQ VYSMGGPNRF 1099
4DA4:A|PDBID|CHAIN| SYHTDINMLY WSDEEAVVNF SDVQGRCTVE YGEDLLESIQ DYSQGGPDRF 357
Consensus SYHXDINXLY WSDEEAVVXF XXVQGRCTVE YGEDLXEXXQ XYSXGGPXRF
Conservation
hDNMT1 NP_001124295.1 YFLEAYNAKS KSFEDPPNHA RSPGNKGKGK GKGKGKGKPK SQ ACEPSEPEIE 1149
4DA4:A|PDBID|CHAIN| YFLEAYNSKT KNFEDPPNHA RSPGNKGKGK GKGKGKGKHQ VSEPKEPEAA 407
Consensus YFLEAYNXKX KXFEDPPNHA RSPGNKGKGK GKGKGKXKXQ XXEPXEPEXX
Conservation
hDNMT1 NP_001124295.1 IKLPKLRTLD VFSGCGGLSE GFHQAGISDT LWAIEMWDPA AQAFRLNNPG 1199
4DA4:A|PDBID|CHAIN| IKLPKLRTLD VFSGCGGLSE GFHQAGISET LWAIEMWDPA AQAFRLNNPG 457
Consensus IKLPKLRTLD VFSGCGGLSE GFHQAGISXT LWAIEMWDPA AQAFRLNNPG
Conservation


1,220
1,240
hDNMT1 NP_001124295.1 STVFTEDCNI LLKLVMAGET TNSRGQRLPQ KGDVEMLCGG PPCQGFSGMN 1249
4DA4:A|PDBID|CHAIN| TTVFTEDCNV LLKLVMAGEV TNSLGQRLPQ KGDVEMLCGG PPCQGFSGMN 507
Consensus XTVFTEDCNX LLKLVMAGEX TNSXGQRLPQ KGDVEMLCGG PPCQGFSGMN
100%
Conservation
0%
1,260
1,280
1,300
hDNMT1 NP_001124295.1 RFNSRTYSKF KNSLVVSFLS YCDYYRPRFF LLENVRNFVS FKRSMVLKLT 1299
4DA4:A|PDBID|CHAIN| RFNSRTYSKF KNSLVVSFLS YCDYYRPRFF LLENVRNFVS YRRSMVLKLT 557
Consensus RFNSRTYSKF KNSLVVSFLS YCDYYRPRFF LLENVRNFVS XXRSMVLKLT
100%
Conservation
0%
1,320
1,340
hDNMT1 NP_001124295.1 LRCLVRMGYQ CTFGVLQAGQ YGVAQTRRRA IILAAAPGEK LPLFPEPLHV 1349
4DA4:A|PDBID|CHAIN| LRCLVRMGYQ CTFGVLQAGQ YGVAQTRRRA IILAAAPGEK LPLFPEPLHV 607
Consensus LRCLVRMGYQ CTFGVLQAGQ YGVAQTRRRA IILAAAPGEK LPLFPEPLHV
100%
Conservation
0%
1,360
1,380
1,400
hDNMT1 NP_001124295.1 FAPRACQLSV VVDDKKFVSN ITRLSSGPFR TITVRDTMSD LPEVRNGASA 1399
4DA4:A|PDBID|CHAIN| FAPRACQLSV VVDDKKFVSN ITRLSSGPFR TITVRDTMSD LPEIQNGASN 657
Consensus FAPRACQLSV VVDDKKFVSN ITRLSSGPFR TITVRDTMSD LPEXXNGASX
100%
Conservation
0%
1,420
1,440
hDNMT1 NP_001124295.1 LEISYNGEPQ SWFQRQLRGA QYQPILRDHI CKDMSALVAA RMRHIPLAPG 1449
4DA4:A|PDBID|CHAIN| SEIPYNGEPL SWFQRQLRGS HYQPILRDHI CKDMSPLVAA RMRHIPLFPG 707
Consensus XEIXYNGEPX SWFQRQLRGX XYQPILRDHI CKDMSXLVAA RMRHIPLXPG
100%
Conservation
0%
1,460
1,480
1,500
hDNMT1 NP_001124295.1 SDWRDLPNIE VRLSDGTMAR KLRYTHHDRK NGRSSSGALR GVCSCVEAGK 1499
4DA4:A|PDBID|CHAIN| SDWRDLPNIQ VRLGDGVIAH KLQYTFHDVK NGYSSTGALR GVCSCAE-GK 756
Consensus SDWRDLPNIX VRLXDGXXAX KLXYTXHDXK NGXSSXGALR GVCSCXEAGK
100%
Conservation
0%
1,520
1,540
hDNMT1 NP_001124295.1 ACDPAARQFN TLIPWCLPHT GNRHNHWAGL YGRLEWDGFF STTVTNPEPM 1549
4DA4:A|PDBID|CHAIN| ACDPESRQFS TLIPWCLPHT GNRHNHWAGL YGRLEWDGFF STTVTNPEPM 806
Consensus ACDPXXRQFX TLIPWCLPHT GNRHNHWAGL YGRLEWDGFF STTVTNPEPM
100%
Conservation
0%
1,560
1,580
1,600
hDNMT1 NP_001124295.1 GKQGRVLHPE QHRVVSVREC ARSQGFPDTY RLFGNILDKH RQVGNAVPPP 1599
4DA4:A|PDBID|CHAIN| GKQGRVLHPE QHRVVSVREC ARSQGFPDSY RFFGNILDRH RQVGNAVPPP 856
Consensus GKQGRVLHPE QHRVVSVREC ARSQGFPDXY RXFGNILDXH RQVGNAVPPP
100%
Conservation
0%
1,620
hDNMT1 NP_001124295.1 LAKAIGLEIK LCMLAKARES ASAKIKEEEA AKD 1632
4DA4:A|PDBID|CHAIN| LAKAIGLEIK LCLLSSA--- ---------- --- 873
Consensus LAKAIGLEIK LCXLXXARES ASAKIKEEEA AKD
100%
Conservation
0%

A3. Chapter IV: Scripts and code used to analyze RRBS and oxo-RRBS data in Chapter IV

A3.1. Generating scripts for primary sample analysis

Using Perl, the first script is called METH_SHOT.pl. The Perl script will write job scripts for each sample that can then perform the following transformations: trimming the adapters, removing the diversity sequences from the sequencing adapters, mapping the bisulfite reads with Bismark, converting the mapped output to a methylated reads input, calculate the bisulfite conversion rate, and calculate the methylated sites. To use the script type “perl METH_SHOT.pl” in the command line, in the same location, you also must have a file “Indexing.txt”. The information contained in Indexing.txt are as follows: column 1 contains the path to locate your fastq files if they are not in the working directory, column 2 contains the name of the R1 file, column 3 contains the name of the R3 file, and column 4 contains the short name which will be the output of the mapping process. The output of the Perl script will be a series of scripts named “scriptA#.pbs”, where # is a number between 1 and the number of rows included in “Indexing.txt”. These PBS scripts can be run using the following command in the command line, “qsub scriptA#.pbs”.

Code for METH_SHOT.pl:

```
$n=1;
open SOURCE, "< Indexing.txt";
while (defined($line = <SOURCE>)) {
   chomp $line;
   print "$line\t";
   @field= split(/\t/, $line);
```

```
   open OUT, "> scriptA$n.pbs";
   print OUT "

#!/bin/bash -l
#PBS -l nodes=1:ppn=24,mem=62GB,walltime=4:00:00
#PBS -m abe
#PBS -M seile053@umn.edu
module load cutadapt/1.8.1
module load fastqc/0.11.7
cd /home/tretyako/shared/Project_006/Analysis/2018-June-01
/home/tretyako/shared/TrimGalore/TrimGalore-0.4.5/trim_galore --paired -a
AGATCGGGAAGAGC -a2 AAATCAAAAAAAC $field[0]$field[1].fastq
$field[0]$field[2].fastq
python /home/tretyako/shared/PythonScripts/trimRRBSdiversityAdaptCustomers.py -1
$field[1]_val_1.fq -2 $field[2]_val_2.fq
module unload cutadapt/1.8.1
module unload fastqc/0.11.7
module load gcc/4.7.2
module load bowtie2
#Test that we can understand MethPipe math with defined data sets
/home/tretyako/shared/Bismark/Bismark_v0.19.0/bismark --multicore 4 --bowtie2
/home/tretyako/shared/Genomes/Mouse-MM10/ -1 $field[1]_val_1_trimmed.fq -2
$field[2]_val_2_trimmed.fq
module unload gcc/4.7.2
module unload bowtie2
#begin conversion to mr files
module load gcc/7.2.0
#GCC version 7.2.0 allows proper calculations.
module load gsl/2.3
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/to-mr -o $field[3].mr -m
bismark $field[4] -v
#Take the mr and sort the file
sort -k1,1 -k2,2n -k3,3n -k6,6 $field[3].mr -o $field[3].t.mr
#Estimate methylation levels
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/methcounts -o
$field[3].meth -c /home/tretyako/shared/Genomes/Mouse-MM10/ -v $field[3].t.mr
#calculate bisulfite conversion levels - these do not need to be 100%, but instead uniform
across samples.
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/bsrate -o $field[3].bsrate -
c /home/tretyako/shared/Genomes/Mouse-MM10/ -v $field[3].t.mr
#END";
```

```
$n++;
}
```

A3.2. Calculating hmC from BS and oxoBS data

The second script uses the BS data and oxoBS data to calculate the levels of hmC at CpG sites. This script is titled “Part2-MLML.pl” and can be executed by typing “perl Part2-MLML.pl” into the command line. Executing the script requires a text file titled “Index-hmC.txt” in the same location. The contents of the text file include a column of the methylated files, the bisulfite sequencing files, and a column of the desired output names.

Code for Part2-MLML.pl:

```
$n=1;
open SOURCE, "< Index-hmC.txt";
while (defined($line = <SOURCE>)) {
   chomp $line;
   print "$line\t";
   @field= split(/\t/, $line);

   open OUT, "> scriptB$n.pbs";
   print OUT "

#!/bin/bash -l
#PBS -l nodes=1:ppn=12,mem=32GB,walltime=12:00:00
#PBS -m abe
#PBS -M seile053@umn.edu
#All of the information above is needed to properly run the commands using the PBS, they are NOT treated as comments.
module load gcc/7.2.0
#GCC version 7.2.0 allows proper calculations.
module load gsl/2.3
#subtract the MeC data in oxBS sample from the BS sample. Gives hmC result
cd /home/tretyako/shared/Project_006/Analysis/2018-June-01
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/mlml -u $field[0] -m $field[1] -o $field[2] -H $field[2].counts
#END";
$n++;
```

}

A3.3. Determining sites with statistically different methylation and hydroxymethylation

This last code is contained in "METHFILTER.sh" and contains data filters and statistical calculations. Initially the various data files containing MeC or hmC are merged into a single file containing the coverage and methylated reads of each sample. These data are then filtered with a read cut-off of 200 reads across all samples to remove regions with little or no data because of RRBS preparation of samples. Differences in methylation are estimated using the FoldChangeFilter-05.pl which filters out sites that have less than a 5% difference between treated and control samples. The statistical differences were calculated using radmeth, and sites with a p-value less than 0.001 were moved to a new file to be included.

Code for METHFILTER.sh:

```
#!/bin/bash -l
#PBS -l nodes=1:ppn=4,mem=5GB,walltime=24:00:00
#PBS -m abe
#PBS -M seile053@umn.edu
#Merge the *.meth files together
cd /home/tretyako/shared/Project_006/Analysis/2018-June-01
module load gcc/7.2.0
#Merge the multiple MethCount files together
#Duration: 5-6 hours
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/merge-methcounts -t
oxCtrl1.meth oxCtrl2.meth oxCtrl3.meth oxCtrl4.meth oxECS1.meth oxECS2.meth
oxECS3.meth oxECS4.meth > MeCData.txt
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/merge-methcounts -t
hmCtrl1.meth.counts hmCtrl2.meth.counts hmCtrl3.meth.counts hmCtrl4.meth.counts
hmECS1.meth.counts hmECS2.meth.counts hmECS3.meth.counts hmECS4.meth.counts
> hmCData.txt
```

```
#
#Filter the merged files to ensure there is enough data for further comparisons and reduce statistical tests
#Duration: 2 hours
#Move the descriptor of the file to the new file and then perform the sort.
head -n 1 MeCData.txt > MeCData_400.txt && awk '{sum=0; for(i=2; i<=NF; i++) sum += $i; if (sum>400) print}' MeCData.txt >> MeCData_400.txt
head -n 1 hmCData.txt > hmCData_500.txt && awk '{sum=0; for(i=2; i<=NF; i++) sum += $i; if (sum>500) print}' hmCData.txt >> hmCData_500.txt
#
#Filter based on estimated fold change of at least 5%. Adjust this value according to depth and needs. Perl Script written by John Garbe, UMGC.
#It is entirely possible that hmC and MeC need to have separate FC filters.
perl /home/tretyako/shared/Project_006/FoldChangeFilter-05.pl MeCData_200.txt > MeCData_200_FC05.txt
perl /home/tretyako/shared/Project_006/FoldChangeFilter-05.pl hmCData_200.txt > hmCData_200_FC05.txt
#
#After filters, statistical changes can be calculated. These have not been adjusted yet.
#Use the header of the file to crease "Design_Matrix.txt" and "Design_Matrix2.txt"
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/radmeth regression -factor case Design_Matrix.txt MeCData_200_FC05.txt > 200_FC05-MeC_CpGs.bed
/home/tretyako/shared/MethPipeSoftware/methpipe-master/bin/radmeth regression -factor case hDesign_Matrix.txt hmCData_200_FC05.txt > 200_FC05-hmC_CpGs.bed
#
#Filter the regression data.
#First on values less than 0.001. Note the line stops reading at the 1, and the values included are actually less than 0.002.
awk '$5 <= 0.001 "{print $0; $}"' 200_FC05-MeC_CpGs.bed > 200_FC05-MeCintmed.bed
awk '$5 <= 0.001 "{print $0; $}"' 200_FC05-hmC_CpGs.bed > 200_FC05-hmCintmed.bed
#Remove rows where statistics could not be calculated. We expect this to be few due to the earlier perl filtering.
awk '$5 > -1 "{print $0; $}"' 200_FC05-MeCintmed.bed > 200_FC05-MeCintmed2.bed
awk '$5 > -1 "{print $0; $}"' 200_FC05-hmCintmed.bed > 200_FC05-hmCintmed2.bed
#Sort the output values based on the given p-value.
sort -k 5n 200_FC05-MeCintmed2.bed > 200_FC05-MeCoutput.bed
sort -k 5n 200_FC05-hmCintmed2.bed > 200_FC05-hmCoutput.bed
#This list should be short enough to export into Excel and get the gene name.
#END
```

Code for FoldChangeFilter.py:

```
#!/usr/bin/perl -w
die "USAGE: FoldChangeFilter.pl inputfile.txt\n" unless ($#ARGV == 0);
# Assign the columns of data to groups for comparing the fold change
%groups = (
   2 => 1,
   3 => 2,
   4 => 1,
   5 => 2,
   6 => 1,
   7 => 2,
   8 => 1,
   9 => 2,
   10 => 3,
   11 => 4,
   12 => 3,
   13 => 4,
   14 => 3,
   15 => 4,
   16 => 3,
   17 => 4,
);
# open input file
$ifile = shift @ARGV;
open IFILE, "$ifile" or die "Cannot open $ifile: $!\n";
# first line is a header, grab it and print out
$header = <IFILE>;
print $header;
# go through the rest of the input file line by line
while ($line = <IFILE>) {
   chomp $line; # remove newline character
   next if ($line =~ /^#/); # ignore lines starting with #
   next if ($line =~ /^\s*$/); # ignore empty lines
   @line = split /\t/, $line; # split line by tabs
   # Go through each column, calculate sum of each group
   @sum = (0, 0, 0, 0, 0); # empty the sum array
   for $index (0..$#line) {
         $column = $index + 1;
         if (defined($groups{$column})) { # if a group is defined for this column
            $group = $groups{$column};
            $sum[$group] += $line[$index]; # add to the group sum
```

```
        }
    }
    # calculate foldchange
    if (($sum[1] + $sum[2] == 0) or ($sum[3] + $sum[4] == 0) or ($sum[4] == 0)) {
       $fc = 1.0
    }
    else {
    $fc = ($sum[2] / ($sum[1])) / ($sum[4] / ($sum[3]));
    $fc = abs($fc);
    }
    # print out the line if foldchange is greater than cutoff. As in it increased across the
groups.
    #Currently 10% change in either direction as below.
    if ($fc > 1.05) {
          print "$line\n";
    }
    # reverse the equation in case methylation or hmC is lost in these sites.
    if (($sum[2] == 0) or ($sum[3] + $sum[4] == 0)) {
       $fc = 1.0
    }
    else {
    $fc = ($sum[4] / ($sum[3])) / ($sum[2] / ($sum[1])) ;
    $fc = abs($fc);
    }
    # print out the line if foldchange is less than a cutoff. This would show a decrease.
    if ($fc > 1.05) {
    print "$line\n";
    }
}
#END
```

A4. Chapter VI: Affinity Proteomics Protocols

Biotin Capture: Reference: "Dynamic Readers of 5-(hydroxy)methylcytosine and its oxidized derivatives" Cell 2013.

Protein Binding Buffer (PBB): 150 mM NaCl, 50 mM Tris-HCl, 1 mM DTT, 0.0025% NP40, pH 8.0.

DNA Binding Buffer (DBB): 1M NaCl, 10 mM Tris-HCl, 1mM EDTA, 0.0005% NP40, pH 8.0.

1. Resuspend the Streptavidin coated beads (Sepharose from GE healthcare).
    a. **Note:** To ensure bead homogeneity, mix the vial thoroughly by repeated inversion, gentle vortexing or using a rotating platform.
2. Prepare an aliquot of beads with 30 µL per sample.
    a. Wash beads with 3x600N µL DBB.
        i. Where N = the total number of samples.
    b. Resuspend in 700N µL DBB and split into two equal aliquots.
        i. 1 aliquot for clearing non-specific binders and biotinylated molecules.
        ii. 1 aliquot for affinity purification.
3. Clear non-specific binders from the samples:
    a. **Note**: Perform step 4 at the same time as the steps below.
    b. Separate one of the above aliquots (step 2b) from DBB and resuspend in 600N µL PBB.

c. Add protein (800N µg, where n is the number of samples) to the beads in PBB.

d. Incubate 2 hours at 4°C (cold room) while inverting samples with the Labroller.

e. Separate beads from solution by centrifugation.

f. Aliquot the supernatant to new tubes for affinity purification.

4. For affinity purification:

    a. Add 6.5N nmol of DNA to the second aliquot of beads from step 2b.

    b. Incubate for 2 hours at 4 ℃ with inversion.

    c. Separate the beads from solution by centrifugation, discard DBB.

    d. Resuspend in 300N µL PBB.

    e. Transfer 300 µL resuspended beads to each sample from step 3f above.

    f. Incubate 2 hours at 4°C (cold room) while inverting samples with the Labroller.

    g. Wash beads from each sample 3x with 300 µLPBB, discarding the supernatant each time.

5. Protein elution:

    a. Incubate beads from each sample in 0.1% SDS at 95°C for 10 minutes to disrupt protein-DNA interaction and elute proteins.

    b. Separate beads from solution by centrifugation and transfer supernatant to a pre-washed FASP filter as indicated below.

FASP Procedure (Using 50 mM HEPES pH: 8)

1. Wash each FASP filter with 2 x 200 µL 50:50 water:ACN with 0.1% formic acid, followed by 2 x 200 µL wash with 0.1% formic acid in water.
2. The supernatant (Step 5b) was transferred to an Amicon 3K filter and the salts were washed away by centrifugation at 14,000 x g for 10 minutes.
3. Proteins were washed with 3x 100 µL of 50 mM HEPES to remove binding buffer and centrifuged at 14,000 x g for 10 minutes.
4. The proteins were resuspended in 50 mM HEPES to a final volume of 100 µL
5. **Reduction:** To the solution, 1 µL of the 200 mM DTT/0.1% SDS in 50 mM HEPES was added and the sample was incubated at 55 °C for 1 hour.
6. Immediately before use, 9 mg of iodoacetamide was dissolved with 132 µL of water to make 375 mM iodoacetamide.
7. **Alkylation:** To each solution, 1 µL of the 375 mM iodoacetamide was added and then incubated for 30 minutes at room temperature protected from light.
8. Each tube was centrifuged at 14,000 x g for 10 minutes to remove excess reagent.
9. Samples were washed 3x100 µL 50 mM HEPES with above centrifugation.
10. The samples were suspended in 50 µL 50 mM HEPES and incubated with 1 µg trypsin overnight at 37 ºC in a water bath.

TMT Labeling of tryptic peptides

1. The tryptic peptides were recovered by centrifugation at 14,000 x g for 10 minutes into a clean tube.

2. The Amicon 3K filters had 3x 50 µL 0.1% formic acid added and centrifuged at 14,000 x g for 10 minutes to wash peptides through the filter.
3. One additional wash with 0.1% formic acid in 70% ACN was performed followed by centrifugation at 14,000 x g for 10 minutes.
4. The collected washes for each sample were combined in a glass MS vial with insert, then concentrated to dryness under vacuum to remove the formic acid.
5. Peptides were cleaned-up using C18 spin columns:
    i. Note: maximum loading of 30 µg per column.
    b. Each spin column was washed with 2x 200 µL 50% ACN to activate the resin.
    c. Columns were washed 2x 200 µL of 5% ACN/0.5% TFA to equilibrate the column.
    d. Reconstitute sample in 100 µL 5% ACN/0.5% TFA.
        i. Ensure sample acidity by checking pH is less than 3.
    e. Load sample on spin column and centrifuge to bind
        i. Collect Flow through and reload the flow through for more complete binding.
    f. Wash column with 2x 200 µL 5% ACN/0.5% TFA.
        i. Use 1-2 additional washes if samples contained large amounts of salts such as urea.
    g. Rinse collection tube with 0.1% FA in ACN to remove free plasticizers.

h. Elute peptides into the collection tube with 3x 20 µL 70% ACN.

i. Dry the sample to remove ACN.

6. Once dry, each sample was resuspended in 100 µL of water and quantified using Pierce BCA colorimetric peptide kit according to the manufacturer's instructions.
7. 2.5-10 µg of peptides for each sample was aliquoted to a new glass MS vial for TMT labeling and dried.
    a. If needed, a reference sample can be prepared by adding an equal amount of peptides from each sample to a single MS vial.
        i. Example: If 2 runs are needed to analyze all of the samples, then the total amount of peptide needed for the reference sample would be: Total Reference peptides = 2*(peptide amount for each sample)
8. Once dry, each sample was reconstituted in 35 µL of 50 mM HEPES pH 8.0.
9. To each solution, 5 µL of a TMT-sixplex (19.5 µg/µL)was added to the peptide solution.
10. To the solution, 10 µL of ACN to yield a final solution of 50 µL, and 30% ACN.
    a. Note: 30% ACN is essential for effective TMT-labeling of peptides.
11. Each solution was incubated at room temperature for 2 hours.
12. Following the incubation, 4 µL of 5% hydroxylamine was added to each solution and incubated for 15 minutes.
    a. Note: This solution needs to be prepared fresh by diluting 50% hydroxylamine solution.

13. TMT-labeled samples were **combined** accordingly.
14. The solutions were concentrated to dryness by speed vac to remove the acetonitrile.
15. TMT-labeled peptides were then fractionated using the Pierce High pH Fractionation kit according the manufacturer's directions for TMT-tagged samples.
    a. Note: The kit has fractionation solutions for unlabeled and for TMT-tagged peptides. It is important to use the most appropriate method for the samples.

A5. Chapter VI: Description of Proteins Listed on Figure 5.4.

The proteins here were identified in Experiment 4, **Table 5.2 in Chapter V of this thesis**. Protein descriptions were summarized from the Uniprot database.[394]

Thioredoxin-TXN

Binding preference for: MeC, hmC, fC, and caC

TXN participates in redox reactions through the reversible oxidation of its active center dithiol to a disulfide and catalyzes dithiol-disulfide exchange reactions, including S-nitrosylation of cysteine residues, for example CASP3.

Protein CMSS1- CMSS1

Binding preference for: MeC, hmC, fC, and caC

CMSS1 function by binding poly(A) RNA.

Pyruvate kinase PKM- PKM

Binding preference for: MeC, hmC, fC, and caC

PKM is a glycolytic enzyme that catalyzes the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) to ADP, generating ATP. Stimulates POU5F1-mediated transcriptional activation.

Protein PRRC2A- PRRC2A

Binding preference for: hmC, fC, and caC

PRRC2A may play a role in the regulation of pre-mRNA splicing and poly(A) RNA binding.

Nuclear autoantigen Sp-100- SP100

Binding preference for: MeC and fC

SP100 is a tumor suppressor protein which together with PML is a major constituent of the PML bodies, a subnuclear organelle involved in a large number of physiological processes including cell growth, differentiation and apoptosis. May also regulate TP53-mediated transcription and through CASP8AP2, regulate FAS-mediated apoptosis.

Filaggrin-2- FLG2

Binding preference for: hmC and fC

FLG2 is a calcium ion binding protein.

Junction plakoglobin – JUP

Binding preference for: hmC

JUP is a common junctional plaque protein, which are needed to arrange and position the cytoskeleton and cells within the tissue.

Cellular nucleic acid-binding protein – CNBP

Binding preference for: hmC, fC, and caC

CNBP is a single-stranded DNA-binding protein, with specificity to the sterol regulatory element (SRE) and is involved in sterol-mediated repression.

Myristoylated alanine-rich C-kinase substrate - MARCKS

Binding preference for: C

MARCKS is the most prominent cellular substrate for protein kinase C. This protein is mostly present in the cytoplasm, and binds calmodulin, actin, and synapsin.

Poly(rC)-binding protein 3 - PCBP3

Binding preference for: C

PCBP3 is a single-stranded nucleic acid binding protein that binds preferentially to cytosine containing DNA.

DAZ-associated protein 1 - DAZAP1

Binding preference for: C

DAZAP1 is a RNA-binding protein, which may be required during spermatogenesis.

Far upstream element-binding protein 1 - FUBP1

Binding preference for: C

FUBP1 regulates MYC expression by binding to a single-stranded far-upstream element (FUSE) upstream of the MYC promoter. May act both as activator and repressor of transcription.

Far upstream element-binding protein 2 – KHSRP

Binding preference for: C

KHSRP binds to the dendritic targeting element and may play a role in mRNA trafficking. KHSRP also mediates exon inclusion in transcripts that are subject to tissue-specific alternative splicing and similar to FUBP1, KHSRP may interact with single-stranded DNA from the far-upstream element (FUSE).

Far upstream element-binding protein 3 - FUBP3

Binding preference for: C

FUBP3 may interact with single-stranded DNA from the far-upstream element (FUSE) and activate gene expression.

Heterogeneous nuclear ribonucleoproteins A2/B1 - HNRNPA2B1

Binding preference for: C

Heterogeneous nuclear ribonucleoprotein (hnRNP) that associates with nascent pre-mRNAs, packaging them into hnRNP particles. The hnRNP particle arrangement on nascent hnRNA is non-random and sequence-dependent and serves to condense and

stabilize the transcripts and minimize tangling and knotting. Packaging plays a role in various processes such as transcription, pre-mRNA processing, RNA nuclear export, subcellular location, mRNA translation and stability of mature mRNAs.[395-397]

Serine-threonine kinase receptor-associated protein – STRAP

Binding preference for: C

The SMN complex plays a catalyst role in the assembly of small nuclear ribonucleoproteins (snRNPs), the building blocks of the spliceosome. STRAP plays a role in the cellular distribution of the SMN complex.

Polymerase delta-interacting protein 3 - POLDIP3

Binding preference for: C, fC, and caC

POLDIP3 is involved in regulation of translation. Is preferentially associated with CBC-bound spliced mRNA-protein complexes during the pioneer round of mRNA translation. Contributes to enhanced translational efficiency of spliced over nonspliced mRNAs. Recruits activated ribosomal protein S6 kinase beta-1 I/RPS6KB1 to newly synthesized mRNA.

Heterogeneous nuclear ribonucleoprotein A1 - HNRNPA1

Binding preference for: C

Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection.

Nucleolysin TIAR - TIAL1

Binding preference for: C

TIAL1 is a RNA-binding protein and possesses nucleolytic activity against cytotoxic lymphocyte target cells.

Replication protein A 70 kDa DNA-binding subunit - RPA1

Binding preference for: C

As part of the heterotrimeric replication protein A complex (RPA/RP-A), binds and stabilizes single-stranded DNA intermediates that form during DNA replication or upon DNA stress. It prevents their reannealing and in parallel, recruits and activates different proteins and complexes involved in DNA metabolism. Thereby RPA1 plays an essential role both in DNA replication and the cellular response to DNA damage.[398]

Replication protein A 32 kDa subunit - RPA2

Binding preference for: C

RPA2 functions as part of the heterotrimeric replication protein A complex (RPA/RP-A) with RPA1 above to facilitate binding of ssDNA and its subsequent repair.

SAP domain-containing ribonucleoprotein – SARNP

Binding preference for: C

SARNP binds both single-stranded and double-stranded DNA with higher affinity for the single-stranded form. In addition, SARNP specifically binds to scaffold/matrix attachment regions of DNA.

Exosome complex component MTR3 - EXOSC6

Binding preference for: C and fC

Non-catalytic component of the RNA exosome complex which has 3'->5' exoribonuclease activity and participates in a multitude of cellular RNA processing and degradation events. In the nucleus, the RNA exosome complex is involved in proper maturation of stable RNA species such as rRNA, snRNA and snoRNA, in the elimination of RNA processing by-products and non-coding 'pervasive' transcripts, such as antisense RNA species and promoter-upstream transcripts, and of mRNAs with processing defects, thereby limiting or excluding their export to the cytoplasm.

Serine/arginine-rich splicing factor 8 - SRSF8

Binding preference for: C

Involved in pre-mRNA alternative splicing.

Bifunctional polynucleotide phosphatase/kinase – PNKP

Binding preference for: fC

Plays a key role in the repair of DNA damage, functioning as part of both the non-homologous end-joining (NHEJ) and base excision repair (BER) pathways. Through its two catalytic activities, PNK ensures that DNA termini are compatible with extension and ligation by either removing 3'-phosphates from, or by phosphorylating 5'-hydroxyl groups on, the ribose sugar of the DNA backbone.

Myosin-9 - MYH9

Binding preference for: fC

Cellular myosin that appears to play a role in cytokinesis, cell shape, and specialized functions such as secretion and capping. During cell spreading, plays an important role in cytoskeleton reorganization, focal contacts formation.

Zinc finger protein 185 - ZNF185

Binding preference for: fC

May be involved in the regulation of cellular proliferation and/or differentiation.

Non-histone chromosomal protein HMG-17 - HMGN2

Binding preference for: hmC, fC, and caC

Binds to nucleosomal DNA altering the interaction between the DNA and the histone octamer. May be involved in the process which maintains transcribable genes in a unique chromatin conformation (By similarity).

High mobility group protein B2 - HMGB2

Binding preference for: fC and caC

In the nucleus, HMGB2 is an abundant chromatin-associated non-histone protein involved in transcription, chromatin remodeling and V(D)J recombination and probably other processes. Binds DNA with a preference to non-canonical DNA structures such as single-stranded DNA. Can bend DNA and enhance DNA flexibility by looping thus providing a mechanism to promote activities on various gene promoters by enhancing transcription factor binding and/or bringing distant regulatory sequences into close proximity. [399-401]

Adipogenesis regulatory factor – ADIRF

Binding preference for: C

Plays a role in fat cell development; promotes adipogenic differentiation and stimulates transcription initiation of master adipogenesis factors like PPARG and CEBPA at early stages of preadipocyte differentiation. Its overexpression confers resistance to the anticancer chemotherapeutic drug cisplatin.

Transcriptional enhancer factor TEF-4 - TEAD2

Binding preference for: C

TEAD2 is a transcription factor which plays a key role in the Hippo signaling pathway, a pathway involved in organ size control and tumor suppression by restricting proliferation and promoting apoptosis. TEAD2 acts by mediating gene expression of YAP1 and WWTR1/TAZ, thereby regulating cell proliferation, migration and epithelial mesenchymal transition (EMT) induction.

14-3-3 protein sigma – SFN

Binding preference for: MeC and hmC

SFN is an adapter protein implicated in the regulation of a large spectrum of both general and specialized signaling pathways. SFN binds to a large number of partners, usually by recognition of a phosphoserine or phosphothreonine motif resulting in the modulation of the activity of the binding partner. When SFN is bound to KRT17, it regulates protein synthesis and epithelial cell growth by stimulating Akt/mTOR pathway.

E3 ubiquitin-protein ligase ZFP91 – ZFP91

Binding preference for: MeC, hmC, and fC

ZFP91 is an atypical E3 ubiquitin-protein ligase that mediates 'Lys-63'-linked ubiquitination of MAP3K14/NIK, leading to stabilized and activated MAP3K14/NIK. It thereby acts as an activator of the non-canonical NF-kappa-B2/NFKB2 pathway.

Non-histone chromosomal protein HMG-14 - HMGN1

Binding preference for: hmC, fC, and caC

HMGN1 binds to the inner side of the nucleosomal DNA thus altering the interaction between the DNA and the histone octamer. May be involved in the process which maintains transcribable genes in a unique chromatin conformation.

Peroxiredoxin-1 - PRDX1

Binding preference for: C and hmC

PRDX1 is involved in redox regulation of the cell. PRDX1 reduces peroxides with reducing equivalents provided through the thioredoxin system but not from glutaredoxin and it may play an important role in eliminating peroxides generated during metabolism.

Annexin A2 - ANXA2

Binding preference for: C, hmC, and fC

Calcium-regulated membrane-binding protein whose affinity for calcium is greatly enhanced by anionic phospholipids. It binds two calcium ions with high affinity.

RNA-binding protein 4 - RBM4

Binding preference for: C

RBM4 in a RNA-binding factor involved in multiple aspects of cellular processes like alternative splicing of pre-mRNA and translation regulation. Modulates alternative 5'-

splice site and exon selection. Exerts a suppressive activity on Cap-dependent translation via binding to CU-rich responsive elements within the 3'UTR of mRNAs, a process increased under stress conditions or during myocytes differentiation. Recruits EIF4A1 to stimulate IRES-dependent translation initiation in response to cellular stress. Associates to internal ribosome entry segment (IRES) in target mRNA species under stress conditions.

Histone H2AX - H2AFX

Binding preference for: C

Variant histone H2A which replaces conventional H2A in a subset of nucleosomes. Nucleosomes wrap and compact DNA into chromatin, limiting DNA accessibility to the cellular machineries which require DNA as a template.